# SOCIAL GAME RETRIEVAL FROM UNSTRUCTURED VIDEOS

A Dissertation
Presented to
The Academic Faculty

by

Ping Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
August 2010

# SOCIAL GAME RETRIEVAL FROM UNSTRUCTURED VIDEOS

Approved by:

Dr. James M. Rehg, Advisor
College of Computing
*Georgia Institute of Technology*

Dr. Thad Starner
College of Computing
*Georgia Institute of Technology*

Dr. Gregory D. Abowd
College of Computing
*Georgia Institute of Technology*

Dr. Rahul Sukthankar
Intel Labs Pittsburgh
Robotics Institute, Carnegie Mellon University

Dr. Aaron Bobick
College of Computing
*Georgia Institute of Technology*

Date Approved: 24 June 2010

*To my parents*

# ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my Ph.D. advisors, Dr. Gregory D. Abowd and Dr. James M. Rehg, for their continuous support and advices through my Ph.D. life. It was their profound vision and immense knowledge that helped me choose the thesis topic, which has generated strong interest in the computer vision community. It was their encouragement and patience that let me endure the tough times with confidence. Thank you very much, Jim and Gregory!

The dissertation is simply impossible without the insights and experience shared by our collaborators from Emory Autism Center, Marcus Autism Center, Children's Healthcare of Atlanta. A special thanks to Dr. Grace T. Baranek at the University of North Carolina at Chapel Hill for her work on early detection of autism that motivated my thesis, and for kindly educating me on various questions that raised during the trip to her laboratory. Special thanks to Dr. Rosa Arriaga and Dr. Agata Rozga for generously offering experts' opinions on autism work. Their inputs have played an important role in the formulation of the thesis topic.

I would also like to thank my committee members: Dr. Aaron Bobick, Dr. Thad Starner and Dr. Rahul Sukthankar. Their insightful comments and suggestions have made me think deeper on the thesis research problems. Thanks to Dr. Irfan Essa who has always been interested in the work and has given many valuable suggestions during the thesis research. Thanks to Dr. Alberto Apostolico for generously educating me on pattern mining during the early phase of the thesis work. Thanks to Dr. Pietro Perona for his intriguing questions. Thanks to Dr. Sing Bing Kang for always offering suggestions on research problems during his visits to Georgia Tech.

Many thanks go to my labmates and friends at Georgia Tech: Matt Mullin,

Howard Zhou, Jay Summet, Matt Flagg, Raffay Hamid, Alton Patrick, Yi Yang, Tracy Westeyn, Mario Romero, Franziska Meier, Mingxuan Sun, Bola Osuntogun, Charlie Brubaker, Gabriel Brostow, Pei Yin, Huamin Wang, Jianxin Wu, Jie Sun, Yushi Jing, Sang Min Oh, Yuting Ye, Yu-Ying Liu, Kai Ni, Kihwan Kim, Matthias Grundmann, Dongshin Kim, Karthir Prabhakar, Wei Cai, Tucker Hermans ... The list is too long to be finished here. I am grateful for all your help with data collection, code sharing, and the exciting discussions. My life at Georgia Tech would not be so enjoyable without you.

My gratitude also goes to the Microsoft Research program in Intelligent Systems for Assisted Cognition, Children's Healthcare of Atlanta, and Google Research for supporting the thesis research.

Finally, I thank my parents for being tremendously supportive and caring during this exciting and stressful time in my life. Also, thank my husband for being a great source of comfort to me in hard times. I can't have come this far without all of you.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# SUMMARY

Parent-child social games, such as peek-a-boo and patty-cake, are a key element of an infant's earliest social interactions. The analysis of children's behaviors in social games based on video recordings provides a means for psychologists to study their social and cognitive development. However, the current practice in the use of video for behavioral research is extremely labor-intensive, involving many hours spent extracting and coding relevant video clips from a large corpus. From the standpoint of computer vision, such real-world video collections pose significant challenges in the automatic analysis of behavior, such as cluttered backgrounds, the effect of varying camera angles, clothing, subject appearance and lighting. These observations motivate my thesis work — automatic retrieval of social games from unstructured videos. The goal of this work is both to help accelerate the research progress in behavioral science and to take the initial steps towards the analysis of natural human interactions in natural settings.

Social games are characterized by repetitions of turn-taking interactions between the parent and the child, with variations that are recognizable by both of them. I developed a computational model for social games that exploits the temporal structure over a long time-scale window as quasi-periodic patterns in a time series. I presented an unsupervised algorithm that mines the quasi-periodic patterns from videos. The algorithm consists of two functional modules: converting image sequences into discrete symbolic sequences and mining quasi-periodic patterns from the symbolic sequences. When this technique is applied to video of social games, the extracted quasi-periodic patterns often correspond to meaningful stages of the games.

The retrieval performance on unstructured, lab-recorded videos and real-world family movies is promising. Building on this work, I developed a new feature extraction algorithm for social game categorization. Given a quasi-periodic pattern representation, my method automatically selects the most relevant space-time interest points to construct the feature representation. Our experiments demonstrate very promising classification performance on social games collected from YouTube. In addition, the method can also be used to categorize TV videos of sports rallies, demonstrating the generality of this approach. In order to support and encourage more research on human behavior analysis in realistic contexts, a video database of realistic child play in natural settings has been collected and is published on our project website (http://www.cc.gatech.edu/cpl/projects/socialgames), along with annotations.

The unsupervised quasi-periodic pattern mining method represents a substantial generalization of conventional periodic motion analysis. Its generality is evaluated by retrieving motions of a range of quasi-periodicity from unstructured videos. The performance was compared with that of a periodic motion detection method based on motion self-similarity. Our method demonstrates superior retrieval performance with a 100% precision when the recall is up to 92.04%, with much fewer parameters than that of the other method.

# CHAPTER I

# INTRODUCTION

## *1.1 Thesis Statement*

Repetition of turn-taking in dyadic social games, as revealed by space-time sequential motion structures, supports effective retrieval of these games from unstructured real-world videos.

## *1.2 Motivation – Video Editing and Behavior Coding in the Early Detection of Autism*

Autism spectrum disorders (ASD) are developmental disorders characterized by impaired social interaction and communication. Currently, the diagnosis of autism is rarely received before age 3 or 4. Early detection of autism has been a growing research emphasis led by the widespread availability of early intervention and the growing evidence of its efficacy for children with autism [60, 71, 81]. In particular, the first randomized, controlled trial conducted by Dawson and her colleagues demonstrated the efficacy of age-appropriate developmental behavioral intervention [20]. Their results showed that effective therapy can lead to raised IQ levels and improved language skills and behavior if a toddler is diagnosed with autism as early as 18 months of age.

The efficacy of early intervention underscores the importance of early detection, as well as early treatment. Currently, both the diagnosis and the treatment of autism are mainly based on behavior assessment in both clinical and natural settings. For diagnosis purpose, the clinician often interacts with the child at clinics according to a protocol, such as the Autism Diagnostic Observation Schedule (ADOS) [46]. For early detection of autism, researchers measure the behaviors both at clinics and in natural settings (such as home and daycare). One popular methodology of finding early signs

of autism is called retrospective video study, by which psychologists examine the behaviors that are found in the child's early life in home movies [2, 7, 16]. These studies indicate that there were indeed symptoms of autism before the child received diagnosis, such as the decreased diversity of social interaction gestures [16].[1] In order to gather uniform objective data about development and characteristics of early interaction patterns, early object exploration and motor patterns, early vocalizations, and sensory responses, prospective longitudinal studies are designed [4]. Longitudinal studies only became feasible when researchers found that siblings of children with ASD had a $3-8\%$ risk of developing autism themselves [3]. The method observes infants from birth to the moment when the diagnosis is received. Prospective studies measure behaviors in both controlled clinical settings and home environments [60, 71, 81]. Interventions, especially preschool intervention programs, are divided into two models: comprehensive and focused interventions [60]. Comprehensive models are broad and usually have multiple components. They can address several developmental domain or skill areas over multiple settings (such as school, clinic or home) or over an entire instructional day. Focused interventions target specific developmental or behavioral outcomes for children. They specify an individual procedure rather than a comprehensive set of procedures, and often are components of a large comprehensive treatment procedures.

The behavior assessment in videos is unavoidable in the detection and treatment of autism. The current practice in behavioral research that involves behavior coding from videos, such as retrospective video studies, are very labor-intensive and time-consuming. Researchers often spend many hours extracting relevant video clips from a large corpus. The behaviors in those video clips are then coded based on frame-by-frame analysis by trained professionals. Sometimes only a small subset of the

[1] Bruner proposes a taxonomy that classifies gestures into three broad categories: social interaction, joint attention and behavior regulation [11].

relevant video clips is analyzed in order to publish research results in time. To reduce the impact of human subjectivity on behavior assessment, interrater reliability is evaluated over multiple independent coders' coding. Such a sheer effort in the video content filtering and behavior quantification is a barrier to research progress. My thesis is motivated by the desire for novel techniques to support efficient video filtering and behavior measurement, in order to reduce the time and labor involved in the current practice, and to enforce consistency of behavior measurement. This work is an initial effort in a new research area led by Georgia Tech that we refer to as **Behavior Imaging**, the purpose of which is to develop integrated technologies for multi-modal computational sensing and modeling to *capture*, *measure*, and *understand* human behaviors.

Parent-child social games are a key element of an infant's earliest social life. Many social interaction gestures, such as hand clapping, reach, wave arms, and shake head no, can be elicited from infants by playing games. Other gestures, such as joint attention, can also be developed through playing games. In addition, social games provide a paradigm for studying many aspects of child development due to the highly structured social patterns that they entail [27].

The goal of this work is to develop *methods for automatically retrieving instances of social games from a corpus of unstructured videos*, such as home movies. From a computational perspective, social games are a natural starting point in addressing the video analysis of social interactions. First, these games typically arise between 9 and 12 months of age, when infants have not yet achieved full mobility, making it easier to monitor the game activity with a single camera. Second, many social games, such as peek-a-boo and so-big, are characterized principally by gross motor movements such as covering and reaching. Although facial expressions encode many subtle cues of the behaviorally-important information, our hypothesis is that it is sufficient to focus on gross movements for the purpose of *distinguishing* social games

from other forms of video content, as is needed in a retrieval task. Third, social games are well structured interactions and individual games typically follow a regular turn-taking pattern. These observations suggest that analysis of repetitive patterns of gross motion in video could be sufficient to identify instances of social games.

From the standpoint of computer vision, the real-world video collections that are used in retrospective video studies pose significant challenges in the automatic analysis of behavior, due to the cluttered backgrounds, the effect of varying camera angles, clothing, subject appearance and lighting. There has been an increasing interest in studying human activities "in the wild." This notion refers to videos of natural activities in natural settings. Examples include the Hollywood action dataset [42] and the YouTube action dataset [45]. In contrast, previous action datasets feature simple actions recorded in a controlled environment, such as the KTH action dataset [75]. We believe that the challenges of extracting behavioral data from home movies is a canonical example of activity analysis in the wild.

In addition to the challenges presented by home movies, social games are well characterized by repetitions of turn-taking between the two players, and are extensively studied by psychologists for various research goals, such as children's social and cognitive development trajectory, how infants grow their social skills [32], parent-infant synchrony [25], screening for problematic parent-infant interactions [26], and the use of games as therapy for developmental delays [35]. These studies of social games provide psychological support and guidance for conducting computer vision based behavior understanding, which in turn may provide more means for psychologists to observe, record and analyze behaviors.

The video dataset we have collected featuring children's play is an important new addition to the existing activities that are studied in the computer vision community. No such complex human interactions have been collected and analyzed so far. Two related video datasets are the Hollywood dataset [42], and the YouTube

action dataset [45]. But those datasets were used for evaluating the capabilities of various features, models and algorithms, rather than behavior understanding. Other activity datasets include American sign languages [83], medical care [78], and kitchen activities [44, 95], which are from a single person's activities.

These observations motivate my thesis work — automatic retrieval of social games from unstructured videos. The goal of this work is both to help accelerate the research progress in behavioral science, and also to take the initial steps towards the analysis of natural human interactions in natural settings.

## *1.3 Contributions*

This thesis research has made four contributions:

1. A new problem to computer vision — analysis of social interactions in unstructured videos. We have published our *ChildPlay* video dataset and annotations to the research community to support and encourage further research on human behavior analysis in realistic contexts.

2. A computational model of social games as quasi-periodic events in a time series, and an unsupervised algorithm that mines the quasi-periodic patterns in videos. This method represents a substantial generalization of conventional periodic motion analysis.

3. A effective categorization of YouTube videos of social games based on bag-of-words model and quasi-periodic pattern mining that selects characteristic visual words automatically.

4. A advocate for more research in order to understand human behaviors, and make fundamental influence to the current diagnosis and treatment of behavioral and developmental disorders.

# CHAPTER II

# SOCIAL GAME RETRIEVAL: PROPERTIES AND CHALLENGES

The goal of this chapter is to set up the background knowledge on social games, and derive the properties and challenges of social game retrieval from unstructured videos. The chapter has three parts. First, we summarize the parent-child social games' characteristics, the diversity among the games, and their role in child development from psychology research literature. Next, we describe an example retrospective video study process used in the early detection of autism to illustrate how video-based behavior assessment is conducted in behavioral science. Finally, we describe the properties and challenges of social game retrieval based on the characteristics of social games and the setup for behavior analysis in natural environments.

## *2.1 Background on Social Games*

Social games, a key element of an infant's earliest social interactions, play an important role in facilitating social and cognitive development [12, 29]. As a result, social games provide a useful mechanism for the study of many aspects of child development, such as social processes and emotion [27], social expectation [70], developmental trajectory [32], development of social skills (*e.g.*, non-verbal communicative skills), sensory-motor skills [7], the adults role in communication with infants [5], and use as a therapy for atypically developed children [35].

### 2.1.1 Characteristics of social games

Social games are two-person (dyadic) interactions governed by an abstract game rule with four main characteristics [28, 34]:

- mutual involvement;
- turn-taking;
- repetition of a two-partner sequence, with a permissable range of variations; and
- nonliterality.

Mutual involvement means that both partners are clearly involved in the interactions, and each acts in response to the other. For example, an infant may throw a ball to his mother, smiling and waiting and looking at her. The mother realizes this is a signal for starting the ball game and rolls the ball back to the infant, and the infant receives the ball, smiles and rolls it back. We say these interactions have become a game. On the other hand, if the mother ignores the signal of playing a ball game, and only receives the ball and smiles at the infant, the interactions are not considered a game. Mutual involvement explicitly requires that both players should contribute to the interaction. Previous research has shown that, on average, typically developing children begin to participate in and initiate social games between 8-12 months of age [17]. However for children with autism, there are inconsistent literatures on whether or not fewer child-initiated gestures is associated with autism status [16, 51, 91].

Turn-taking can be viewed as the expression of mutual involvement. Each partner has his own role to play in the game, and turn alternation regulates the interaction. Successful turn-taking demonstrates the skill of leaving room for the other to act and signaling the other to take his turn. The signals that regulate turn alternation vary widely with games. Typical signals include waiting after completing one's own action, withdrawing physically from the action (such as stepping back, sitting back) and pausing, sometimes accompanied by gaze exchange. Turn-taking is also observed in many other competitive games, such as tennis, table tennis, volleyball (turn-taking

(a) Expecting and reaching.

(b) Surprised at unexpected reappearance of the stuffed monkey toy.

**Figure 1:** Two moments in a peek-a-boo game.

between the two teams). The turn-taking signal can be verbal or nonverbal. Prelinguistic infants usually use nonverbal behaviors or eye-contact, or even simple sounds.

Repetition could make an extended sequence of interaction easier for the infants, as each turn does not need to be completely novel [34]. It may also help the infants to focus on the maintenance of the flow of interaction, rather than the content of each turn. Figure 1 shows two moments of a father playing peek-a-boo with his baby with a stuffed monkey toy. The baby boy reached to the potential position of the monkey with hands (Figure 1(a)) and laughed vigorously every time the toy reappeared in front of him, even when the position and the timing of reappearance differed from the previous turn (where he had expected and reached for). When the father showed the stuffed monkey unexpectedly to the infant's side (Figure 1(b)), still with the signal of sound "Boo", the infant was surprised initially and didn't smile until the father shook the toy closely by his face for a while. The variation in Figure 1(b) was too much for the infant to recognize as part of the game. These two cases illustrate that while the repetitive turn-taking does not need to repeat the previous actions exactly, the variations in turn-taking do need to be accepted by both partners.

Describing a behavior as non-literal is subjective, and is based on an underlying assumption of a literal interpretation of the same behavior. For example, the action

of an infant placing rings scattered on the floor into a container might have a literal explanation as cleaning up the room. However, if after gathering the rings up, the infant throws them onto the floor, and then he collects them and throws them out again, and so on, then we say that infant's behavior is non-literal. The ball game can serve as another example. If the infant receives the ball and then walks away, we may say he interprets the action of receiving the ball literally as "take." If he receives the ball and rolls it back to the mother, with hands reaching forward and a smile on his face, then we say he acts non-literally. These two examples illustrate two facts. First, the definition of literal vs. non-literal is subjective. Second, a single behavior can have multiple context-dependent interpretations. Therefore, assessing an action as non-literal requires inference over supplementary cues such as the antecedent, accompanying and consequent actions of the actor, as well as non-negative or happy emotions being involved. Repetitions, together with all the positive contextual information, are important for identifying non-literal actions. Despite the vague and subjective nature of nonliterality property, it is the unique feature that brings social games to the domain of play [34].

Mutual involvement, turn-taking and repetition describe the interactive nature of the game. They are objective properties that can be observed from the sequence of actions, and they form the basis for our computational model. Nonliterality, on the other hand, is a subjective property inferred from an observer's perspective. The four characteristics together capture a social game's existence from being initiated, to its climax at engaging the infant, and to its ending.

### 2.1.2 Diversity among social games

Infant social games exhibit a wide diversity in the individual acts that they are composed of [34]. There are three sources that contribute to this diversity. First, it is the actions that make up the game. Balls are rolled or thrown; blocks are tapped together

or banged, or used as bowling pins. Infants are able to create activities based on the properties of the physical objects available. The second source of diversity is the extent of variations that are allowed in each turn. An infant may incorporate changes in his own turns and accept changes in the other, which indicate his understanding of the abstract game rule. The third source of diversity is the reversal of roles that occurs during a game. A mother initiates a peek-a-boo game by crouching down, hiding herself from her child; then she stands up to be visible to the child. The child may later request to take on the role of hiding. This suggests that infants can reflect upon the nature of their own role, their partners' role, and the relationships between the two.

### 2.1.3 Functions of social games

Social games provide a rich situation for infants to learn about the properties of the physical world around them. For example, the rings are named by different colors, and they can be stacked onto a cone; the ball can be rolled and be squashed; clapping hands will make sound. Infants also develop behavior regulation skills and motor skills by imitating their partners.

More importantly, social games provide a rich training process for learning interaction skills. Infants learn how to interact, the timing for turns, skills for taking and generating turns, understanding non-literal actions, mastering the regulated and repetitive game structure, creating new variations of each turn, and skills to maintain the continuous flow of interactions.

In the long term, the social interaction skills rehearsed in the games help infants to master conversational interaction, which also has the form of alternating turns [73]. They also help infants to develop new social relationships with peers or strangers, particularly through playing games [34].

## *2.2 Retrospective Video Study*

Retrospective video study is a standard research methodology in behaviorial science [16, 90]. It is a process of collecting and analyzing home movies featuring children with known diagnostic outcomes with the goal of identifying early developmental features of behavioral disorders in a naturalistic setting. It has been used to motivate early detection and intervention of various developmental disorders [13, 35, 7, 16, 90], and validation of regressive autism [89].

One of the primary uses of retrospective video studies is to identify early behavioral traits in children under 2 years old that are associated with a risk for ASD. According to Bruner, gestures are classified into three broad categories: social interaction, joint attention and behavior regulation [11]. Among them, social interaction gestures have received much attention in the work of early detection of ASD for two reasons [7, 16]. First, social interaction gestures are the earliest form of nonverbal communication, and it would potentially provide early signs of autism. Second, there is an increasing need for better understanding of how the usage of dyadic forms of social interaction gestures accounts for the degree of an infant's social-cognitive maturation. Studies have found that children with autism show relative strength in using gestures for behavior regulation, and striking weakness for joint attention [55]. But inconsistent conclusions are made regarding whether or not social interaction gestures are specifically deficient in autism [16, 51, 91]. Colgan *et al.*'s work shows that the decreased variety in type of social interaction gestures is significantly associated with autism status, after a retrospective video study on home movies featuring 9-12-month-old infants with autism and typical development [16]. This is different from previous findings that the number of total gestures and initiation of gestures are significantly associated with autism [91]. As a result, social interaction gestures are of particular interest for the early detection of autism, whose earliest forms are often found in social games.

In this section, we describe a example procedure of retrospective video study used by Baranek and her group at University of North Carolina - Chapel Hill [7, 16]. Other methods in behavioral science that require behavior assessment from videos share similar experience on video content editing, behavior coding, and interrater reliability evaluation.

**Participant recruitment**. The recruitment targeted families who have home videos of their children under 2 years old. They are usually 3 groups of children: typically-developed (TYP group), diagnosed with autism at a later age (AUT group), and diagnosed with other developmental delay at a later age (DD group). The researchers have contacted about 1000 families (by the time of publishing [7]) through personal and professional contacts, advertisements, and direct mailings through hospital-based clinics, public and private schools, early intervention programs, and advocacy groups for child with autism and mental retardation. The overall positive response rate to the mailings was about 10%. Their recruiting process also shows that connections made through personal contacts and professional colleagues were the most successful. As a result, 75 families agreed to share their home videos. After a check of video qualities and content, only 32 families' video tapes (11 AUT, 10 DD, 11 TYP) met their research requirements and were used in the study.

**Participants' contextual information collection**. Contextual information about participants vary with the concrete research goals. Typical information includes infant's age, race, gender, mother's age, father's age, number in family, number of siblings, annual family income, mother's education, father's education [27]. For comparative study across multiple infants groups, such as TYP, DD, and AUT, psychologists also record infant's cognitive score and the time of diagnosis [16].

**Video collection and review procedures**. Copies of the family videos were coded by ID number to preserve confidentiality. The typical scenarios in the videos are family play situations, special events (*e.g.*, birthday parties), vacations, and family

routines (*e.g.*, dinnertime).

The review procedure first marks the scenarios and contents in the videos. Then information on child's chronological age (calculated by full months) during each scenario and specific content was logged. For a particular age range for research purpose, for example, 9- through 12-month age range, the videos were marked for these age groups and were used in later analysis.

**Video editing procedures**. A research assistant who was unaware of the purposes of the study edited the videos. For each subject individually, she was instructed to randomly select a cross-section of scenes from the videos marked with designated age groups. The various scenes were assembled into two 5-minute video segments (A and B). These newly edited video segments were identified by the subject's ID number, followed by A or B. On average, 4 scenes were represented in each 5-minute segment per child. The order of the subjects was randomly mixed onto the final master tapes to be used later for coding purposes.

**Video coding**. The five-minute segment was divided into 20 15-second intervals. Then this data was given to two independent, trained coders who were blind to the purpose of the study for coding. A detailed instruction for coding includes a checklist of behavior categories, and the criteria for these behaviors. The general categories are developed from relevant literature. They include looking and gaze aversion, affect, social touch, postural adjustments, responsiveness to name, motor and object stereotypies, and sensory modulation. Typically, one category is coded at a time.

For each behavior category, measured variables included frequency and scale score. Frequencies were computed across the 20 intervals. Average frequency was computed over the two 5-minute segments, and were used in the final statistical analysis. The scale scores, such as intensity of affective expressions, level of object play, sensory modulation responsiveness/aversion, were quantified using a 4-point rating scale.

**Interrater reliability**. Interrater reliability was obtained by having the raters score approximately 7 to 10 (5-minute) video samples for each behavioral category. A conservative measure of interrater reliability was used by calculating percentage agreement (for positive instances) for each variable per category. Intraclass correlation coefficients (ICC) were also computed for each variable per category. ICC is used to estimate the correlation of one variable between two members within a group.

**Limitations**. The entire data collection and editing process for video coding has two main limitations. One is from the video contents that were provided by parents. Parents tend to preselect the situations that favor pleasant situations and special achievements and avoid videotaping children during uneventful, unpredictable, or adverse conditions – a process that may obscure certain symptoms. Second is from the video editing procedure. It loses social contexts. For example, affective response may need to be measured within specific social contexts or perhaps in tandem with other social responses (*e.g.*, smiling while looking at people) in order to reveal atypicalities at a young age.

As the initial research effort in **Behavior Imaging**, we are interested in creating techniques for the video editing process. The effective video filtering techniques have three benefits to psychologists: prevent selection bias by being able to analyze more video data recorded under more diverse conditions [90]; obtain sufficient data across multiple subjects for fair comparison [7, 16]; and reserve the contextual information to assist behavior assessment (since video editing will not be needed).

## *2.3 Properties and Challenges in Social Game Retrieval*

In this section, we describe the properties of social game from a computational perspective, and summarize the challenges of retrieving social games from unstructured videos.

### 2.3.1 Properties of social game retrieval

For the purpose of retrieval, we are interested in computationally characterizing the actions in social games. The gross motions generated by playing social games have three properties:

- turn-taking interactions;
- repetitions of turn-taking interactions, with a range of variations; and
- multi-instantiation.

In a social game, the parent and the child take turns to act and respond to each other. A turn-taking interaction has two elements: participants of the interaction, and the temporal order of the turns. The basic participants are the parent and the child, which we refer to as a person-to-person interaction. For example, the two people interact directly with each other in a patty-cake game. Sometimes objects are used for the interaction, and we call it a person-object-person interaction. The ball serves as vehicle for communication in a ball game (the interactions occur between actor A, ball and actor B). The temporal order of the turn exchanges can be parallel or sequential. A parallel order means the two partners act simultaneously with similar rhythm. For instance, the two actors push their hands towards each other until hands clap, then withdraw hands together preparing for the next turn. Most social games have a sequential temporal order of alternating turns. In a ball game, releasing the ball must happen before receiving the ball. A peek-a-boo game starts with actor A hiding herself, then A reappears, followed by the elicited laughter from actor B. In reality, there is often an overlapped interval where both turns exist due to anticipation (sometimes referred as expectation), which is similarly found in turn-taking conversation [73]. One would stretch out his hands preparing for catching the ball before the ball leaves the other's hands. The order of the turns can change while

playing, as long as a mutual agreement is established between the parent and the child. A peek-a-boo game may be initiated by the mother hiding herself, and the child may request the role of hiding after several rounds. This is unlike some ordered activities, such as making a French toast, soaking bread (one sub-routine) must happen before cooking the same bread (another sub-routine) [44], or highly-structured medical procedures [78].

The turn-taking interaction is repeated throughout the game, with a range of variations that are acceptable to both actors. A key aspect of the interactions is the *synchrony* that is developed and regulated by both partners. Synchrony (also known as reciprocity) refers to the time ordering of events within the interaction. In a peek-a-boo game, the infant reacts to the (anticipated) reappearance of the parent's face with an expression of joy, and the parent often varies the rhythm of hiding and showing the face in order to keep the child engaged and to avoid overstimulating him. Consequently, the repetition is **quasi-periodic** for two reasons. First, the duration of each turn in an interaction and the overall length of an interaction vary at different repetitions. Second, extraneous actions may be inserted or deleted during a game. Figure 2 shows an inserted action of the mother turning the cloth around between two successive interactions in a peek-a-boo game. In this game, a complete interaction is decomposed into three stages: mother lifts the cloth (Figure 2(a), 2(g)); mother covers the father's face with the cloth (Figure 2(b), 2(h)); mother helps the baby to pull down the cloth to uncover the father's face (Figure 2(c), 2(i)).

Multi-instantiation is a distinct property that results from non-literality and the abstract game rule. The rules determine the content of each player's turn, the allowable variance of each repetition, and the non-literal meaning of the behaviors. We illustrate this property with three example games. A peek-a-boo game has three stages: initial contact, disappearance, reappearance and re-established contact [12]. There are many different ways to play the game, including role-exchanges and a wide

(a) Lift, frame 116 (b) Cover, frame 145 (c) Uncover, frame 278

(d) Frame 403 (e) Frame 437 (f) Frame 450

(g) Lift, frame 462 (h) Cover, frame 481 (i) Uncover, frame 523

**Figure 2:** Between two successive peek-a-boo interactions (top row and bottom row), the mother turned the cloth around (middle row).

(a) Peek-a-boo 1 (b) Peek-a-boo 2 (c) Peek-a-boo 3

**Figure 3:** Three peek-a-boo games from YouTube.

variety of ways to cover and uncover the face, *e.g.*, with a toy, with a cloth, or hiding oneself. Table 1 lists some example variations. Figure 3 shows three peek-a-boo games from YouTube videos. A ball game also has multiple instantiations. The rule of a ball game is that the participants receive and lose the ball in turn. The infant can play passively by "letting" the mother roll the ball to him and then take it away from him. Or, the infant can play actively by rolling the ball back to the mother after receiving it. Patty-cake has three levels of playing from the perspective of the infant's role: participation, imitation and communication depending on infant's skills [1]. In participation level, the mother holds the infant's hands in hers, and pats them with the rhythm. The infant can then play on his own in the imitation level with less help from the mother, where he moves his hands in a similar way to the mother's. In the communication level, the infant actively plays with the mother by clapping the mother's hands.

**Table 1:** Examples of the variations within a peek-a-boo game [12]. M: mother; C: child.

| | |
|---|---|
| Open move | face-to-face mutual looking, vocalization, highlighting an instrument |
| Disappearance pattern | M initiates and M hides C or M; C initiates and C hides C or M. |
| Tools for hiding | clothing, hands, towel, chair, sofa, avert the head |
| Uncovering | M provokes C to uncover, M uncover herself, C uncover M; with vocalizations such as "boo", "Peek-a-boo", "ahhh" |

All social games involve some form of turn-taking and a single game usually consists of multiple repetitions of a basic series of events (that constitute one interaction). As a consequence of these insights we arrive at a fundamental characterization: **social games are defined by quasi-periodic spatio-temporal patterns in videos**. In comparison to a structured activity such as running, social games can exhibit substantial variations in period length and poses. But the existence of a repetitive pattern is a key defining property nonetheless, and it is invariant to the multiple instantiations of the same game.

### 2.3.2 Challenges in social game retrieval

Social games are extended dyadic interactions over a long time scale (several minutes or longer), which is very different from the existing activities that have been studied for decades by computer vision and multimedia retrieval communities. Those activities are limited to single-person activities, such as running and walking [6, 18, 19, 41, 42, 58, 66, 69, 77, 59], or making a sandwich [44, 95], or multi-person interactions over a short time scale, such as hugging, shaking hands [42], and social greetings [62]. The problem of social game retrieval from unstructured videos presents great challenges in both the environments and the activities themselves. To our knowledge, this is the first work of this kind.

We identify two sources of complexities that uniquely arise in the retrieval of social games. The first source is that social games are quasi-periodic. Social games consist of repetitions of a sequential events, with variations. As synchrony is developed during the game, it is inevitably to have varying periods of the entire interactions, and varying durations of individual stages in the interactions. Different poses may be presented as well. For example, in the peek-a-boo game with a stuffed monkey toy (Figure 1), the father may raise the toy to different heights and distances to the child. The second source is from the property of multi-instantiation. It results in a wide range

of patterns of motion and appearance, and multiple versions can be observed within a single game session. Quasi-periodicity and multi-instantiation lead to substantial within class variations, which makes supervised methods very challenging.

Social game retrieval also shares the common challenges with conventional activity recognition, such as the effect of varying camera angles, clothing, subject appearance, and lighting on the captured video. In contrast to visual surveillance applications, we expect social games to be largely in-focus, centered in the frame, and visible due the efforts of the videographer. This imposes constraints on the camera motion which we can leverage to aid in retrieval.

When psychologists conduct behavior coding, they not only look at the action itself, but also evaluate the social aspects accompanied with the action, such as affection, eye-contact, and joint attention. To give automatic fine-grained descriptions of the interactions and to quantify such interactions, one needs to keep track of the gaze, the timing of gaze shift, eye-contact, the evolving of facial expressions, and so on. Automated behavior quantification and understanding are the future research topics in our *Behavior Imaging* research agenda.

# CHAPTER III

# LITERATURE REVIEW ON ACTIVITY RECOGNITION

In this chapter, we review the literature on activity recognition from three perspectives: analysis of activities that exploits activity properties, approaches that explore the representation of activity in videos, and the contextual information for activity recognition. The works that exploit the various activity properties and the works that focus on representing activities in videos often treat the activity of interest independent of their surrounding environments, while the contextual information has been playing a growing important role at recognizing activities in complex scenarios.

## *3.1 Activity Properties*

The temporal structure is an important activity property. Sometimes a key moment pose may be sufficient to characterize the activity, such as the referee signals of football and basketball. Most of the time, it is the sequence of the actions in an activity that tells the full story about the activity.

### 3.1.1 Exploiting repetitiveness

Repetitiveness is a powerful temporal structure cue for video analysis. Many human activities present repetitiveness in time. Periodicity is a strong form of repetitiveness. It is mathematically described as $x(t+T) = x(t)$, where $x(t)$ represents the state at $t$ and $T$ is the constant period. Previous works mainly use Fourier time-frequency methods to analyze videos of periodic motions, such as walking and running [18, 66]. This approach requires either reliable motion segmentation for cluttered background [18], or trajectory estimation of moving objects [66].

In addition to Fourier analysis, periodicity is interpreted as temporal self-similarity

between any two recurrences of a periodic action are similar to each other. Such self-similarity has been used for periodic motion detection and classification [18], periodic motion segmentation [41], and view-invariant action recognition [38]. Laptev *et al.* observed that period-separated frames from a monocular video of a periodic moving object can be approximately viewed as multiple views of the same motion [41]. Their work applied sequence alignment to estimate the period assuming the camera and object are restricted to a constant translation with respect to each other. The autocorrelation of the similarity relationship among all pairs of the frames is shown to present structures that correspond to different types of periodic motions [18]. Junejo *et al.* extended this idea and developed temporal self-similarity based descriptors for action recognition under different views [38]. For image and video matching, local self-similarity descriptors in space (or space-time volume) were proposed by Shechtman and Irani [77]. A pixel descriptor was built by correlating a local patch centered at that pixel with the bigger surrounding area.

Repetitiveness also indicates short recurring motions in very specific scenarios, such as the "rest" moments in a natural story-telling scenario [92], and the primitive actions that form a cashier's checking-out-item sequence at retail stores [23]. The abnormality detection work by Zhong *et al.* is conceptually close to our work [103]. They classify globally-recurring events as normal, and rarely-occurring events as abnormal, with the goal of distinguishing these two classes in surveillance footage. In a similar vein, repetition in space is also used to detect abnormalities [9]. Repetitions of the same short action are used for grouping the motion descriptors extracted from each action's occurrence to discover the atomic movements in human activities [24]. Various motif discovery methods are also developed based on the partial or full repetitiveness of the patterns in multiple sequences [53].

Repetitiveness is used for speech processing as well. Park and Glass's work exploits the temporal structure of repeating patterns within the audio signal stream to extract

the inventory of lexical entities such as words and short, multiword phrases [63]. The extracted repeating entities often are relevant to the underlying lecture audio stream. We plan to incorporate the recurring audio signal (such as "poo" in a peek-a-boo game) to improve the social game retrieval performance in our future work.

A key property of social games is the repetitions of interactions. However, Fourier analysis methods or self-similarities are unlikely to succeed in our task due to the non-constant $T$ in social games, and the generally bigger within class variations in repetitions, compared to the periodic motion (such as walking and running). The repetition in social games requires a weaker form of repetitiveness, which we refer to as *quasi-periodicity*. Our method can extract quasi-periodic sequential motion patterns of any length, and the duration of each occurrence can vary substantially for each pattern [86].

### 3.1.2 Temporal order in activities

Explicit modeling of temporal order in activities allows reasoning in space and time to recognize sequential activities. Stochastic finite state automata (SFSA) are classical models of discrete temporally-evolving phenomena. Causal relationship is a special case of temporal order in activities, has been analyzed a lot in recent activity analysis.

**Explicit temporal relationship modeling** Typical SFSA models include Hidden Markov Models (HMM) [68] and its variations. Encouraged by HMM's success in speech recognition, they have also been used to model American Sign Language gestures [83], gestures for communications [93, 92], even human body configurations in primitive motions (such as walking and running) [69]. To handle individual variations, the parameterized-HMM (PHMM) was developed to explicitly model the personal-dependent systematic spatial variation for gestures, instead of treating such systematic spatial variation as noises [93]. It was also used to handle temporal variations [97]. Coupled-HMM explicitly encodes the interlaced actions from two persons [62]. It can

also model short interaction with objects. The layered-HMM encodes office environmental information, reduces the parameters for plain HMM, and provides more structure for learning [61]. For complex activities, HMM is not efficient to handle partially ordered, parallel sub-routines, since the required state space could grow exponentially.

SFSA in principle can be constructed to model the social game interactions. However, the learning of structure and parameters of a SFSA model needs a prohibitive amount of labeled training data in order to characterize the tremendous variety of possible mappings from a single action state such as "roll the ball" to the space of all possible video clips of that event, due to the multi-instantiation property and various camera views in home videos. In contrast, our quasi-periodic model captures the temporal structure of social games without explicit activity modeling and training, and it is invariant to the instantiation of the game rule.

Dynamic Bayesian networks (DBN) is a related class of techniques to model activities. DBNs are flexible to incorporate various constraints, constituent events to form more complex activity model than HMMs [65, 78, 44]. Pinhanez and Bobick proposed to use the notion of "past", "now", and "future" to model the relative temporal relationship among the events for activity recognition [65]. Temporal, contextual and ordering constraints were modeled by a DBN for recognizing complex activities such as making a sandwich in video [44].

Several previous works have addressed the problem of learning models of highly-structured behaviors from sensor data. Some representative examples are [57], in which the protocols of card games are learned through logical induction, and [36], which addressed Partially Observable Markov Decision Process (POMDP) based learning of decision theoretic behavior models. In comparison to these works, we address a broader class of behaviors, with a focus on retrieval instead of modeling.

**Causality analysis** The dyadic interactions in games may also be characterized

by the causality relationship, *i.e.*, one's action is a response to the other's. Causal relationships are ubiquitous in interactions among objects and people. Newtonian mechanics, an example of the force-effect causal relationship, is used to analyze observations of interacting objects from video [49]. Force-dynamic interpretations are derived from the objects' positions, shapes, velocities *etc.*that are acquired by tracking. Physical-world properties are also used by Brand to analyze the possible motions of objects [10].

In addition to the Newtonian-type relationships in the physical-world, other types of causality can be defined for human actions. For example, the LabelMe video work demonstrated that simple statistics on annotated moving objects in videos could be used to infer causal relationships [101]. Causal relationships between actions can be used to analyze the storyline of activities. In the work of Gupta *et al.*, causal relationships are represented by spatial distance constraints and the temporal ordering of two actions, and is learned through structural EM-like approach [31]. The continuous Granger causal test is used by Zhou *et al.* for pair-activity classification [104]. In their work, trajectories of moving objects were first extracted by tracking, then causality analysis was conducted on the bi-trajectories of two moving objects from the same time window. Supervised methods were then used for the classification task. Loy *et al.* models time-delayed dependencies between spatio-temporal activity patterns explicitly to characterize their causal relationships [48]. The model of these relationships are shown to be able to detect abnormal behaviors effectively. Instead of analyzing causality on extracted trajectories or atomic actions, our recent work [67] applies Granger causal analysis to the visual words extracted from videos in an unsupervised manner. It can automatically identify individual causal groups of visual words and demonstrated improved performance on social game retrieval on our ChildPlay dataset.[1]

[1]The dataset is described in Chapter 4, Section 4.5

## *3.2 Activity Representation*

Instead of decomposing a complex activity into constituent short-term actions, and modeling the temporal relationships among the actions explicitly, there are methods that use the statistical features extracted from a space-time volume to characterize actions, or treat an action as a shape in the 3D space-time space. Combined with learning methods, these representations can give high recognition rate on short, simple actions, or promising results on real-world videos.

### 3.2.1 Sparse Visual Words Representation of Activities

Motivated by the success of sparse features in object recognition [47], sparse space-time features have shown promising performance in recognizing actions [43, 41, 42, 58, 37, 21, 75, 102, 94]. They are tolerant to unrestricted scenarios and large within-class variations due to difference in physical surroundings and human postures [43, 42]. Currently, space-time interest points are mainly applied to recognize short-term primitive actions, or short-term interactions, based on a large corpus of labeled training examples [42]. Enforcing the temporal ordering [59] or the structural information [94] of visual words is shown to increase the discriminative power of the sparse representation.

We build our work on the sparse feature based action representation [40]. Our work suggests, perhaps surprisingly, that meaningful analysis of complex interactions can be obtained by considering the temporal structure of the visual words, without a decomposition into specific actions. In comparison to the work of recognition of human actions in movies [42], our work is different for two reasons: 1) social games are long-term activities consisting of short-term actions; and 2) it is difficult to recognize a social game by a single snapshot of the activity, as many composite actions from different games may look similar. Therefore, even with perfect action detections, quasi-periodic sequential motion pattern extraction is still needed to retrieve social

games from videos.

### 3.2.2 Template Representation of Actions

By constructing 2D motion-history image [19] or 3D shape/template [8, 39, 99, 77] models, activity recognition is formulated as a template matching problem. These approaches are simple and can be made robust to the temporal variations of actions being performed. Most of them require a good separation of the foreground and background. Similarity based on volume intersection reduces the requirement on good foreground segmentation [39]. Its matching between a hand-labeled 3D activity shape and over-segmented super-voxels, with complementary flow matching [76] can detect short-term motion in crowded, dynamic backgrounds. Spatial relationships among the many local feature patches (small templates) were encoded in template matching to increase robustness [9, 77]. Local self-similarities are exploited in template matching to handle complex backgrounds [77]. This approach extracts dense local self-similarity descriptors, and then builds ensembles from local descriptors for both template and testing images/videos. It is robust to cluttered background, lighting change, but it is only valid for short, primitive motions. A substantial obstacle to the use of template matching for social game retrieval is the vast number of templates that would be needed to cover the space of possible games.

### 3.2.3 Modelling Kinematic Structure

Different motions will produce different trajectories of body parts [69, 93, 97] or joint positions [6, 82] at every frame. Body parts-based action recognition often relies on human body tracking and suffers from self-occlusions, dynamic, cluttered background and lighting changes [14, 80, 87]. It is difficult to apply a sequential kinematic structures to model long-term complex activities.

Under constrained environment where the clinicians conduct structured interactions with the children, it is possible to recover the body pose with certain confidence.

In this case, the human poses, and the associated eye gaze and facial expression can be used together to quantize various aspects of social interactions, such as affection and responsiveness. This is an ongoing work in our **Behavior Imaging** research agenda.

## *3.3 Incorporating Contextual Knowledge*

Actions and the objects being used by the actions provide contextual information and can be used to recognize both [54]. The contexts are specified in terms of (object) positions, activity models, and scene layout. The applicable environment should be constrained (*e.g.*, office), with many activities of handling objects, and some objects have fixed positions known in advance.

Object-based activity recognition has received much attention recently. It is motivated by two facts: everyday activity sensing is feasible with more and more non-intrusive, affordable sensors being developed (*e.g.*, RFID); a group of activities are well characterized by the object series being used, such as a procedure for preparing dinner in the kitchen [44, 95]. Non-visual sensor reduces the burden on the vision task and sensor infusion improves recognition rate [95]. Knowledge of object-action relationship are mined from the web [96]. Hierarchical shrinkage and mining ontology from WordNet help researchers maximizing the completeness of such knowledge [56]. Vision-based activity recognition also benefits from making use of the prior knowledge on object-action dictionary [44]. Recognition of objects and activities mutually benefit from each other and are united.

Psychologists report that contextual information sometimes is important for them to recognize the gestures that are otherwise similar. For example, attentional gesture and requesting gesture are both gestures indicating communication. Attentional gestures typically involve looking with head movements, eye gaze, and body stance and position [79]. Requesting gestures also have similar content. Therefore the context in

which those actions are being observed provides crucial information at differentiating these two gestures.

# CHAPTER IV

# QUASI-PERIODIC PATTERN MINING FOR SOCIAL GAME RETRIEVAL

In Chapter 2, Section 2.3, we defined social games as quasi-periodic (QP) spatio-temporal patterns in videos. In this chapter, we describe an unsupervised method that mines such quasi-periodic patterns from videos. We make a key assumption about the manner in which social games are filmed. We expect both actors and any objects, such as balls, to be largely in-focus, centered in the frame, and continuously visible throughout the sequence due to the efforts of the videographer. In addition, we assume the videographers will do their best to hold the camera still throughout the game. In order to detect the quasi-periodic patterns, which indicates the existence of social games, we need to answer two questions: how to extract motions, how to represent and mine the quasi-periodic sequential motions.

Our approach leverages on two facts about social games in videos. First, the social games are characterized principally by gross motor movements, such as covering and reaching. These types of motions often involve the reversal in the direction of movements, or a sudden change in the speed of the motion. Examples include clapping hands in a patty-cake game, the appearance of face followed by a pause in a peek-a-boo game. Such motions generate strong corners in the 3D space-time volume. We use Laptev's space-time interest point detector [40] to detect such characteristic motions. This is particularly important for detecting social games in unstructured videos, as articulated human body detection/tracking will easily fail due to the varying camera angles, camera motions, clothing, lighting, and self-occlusions.

Second, our approach looks at the activities over a long time scale window for

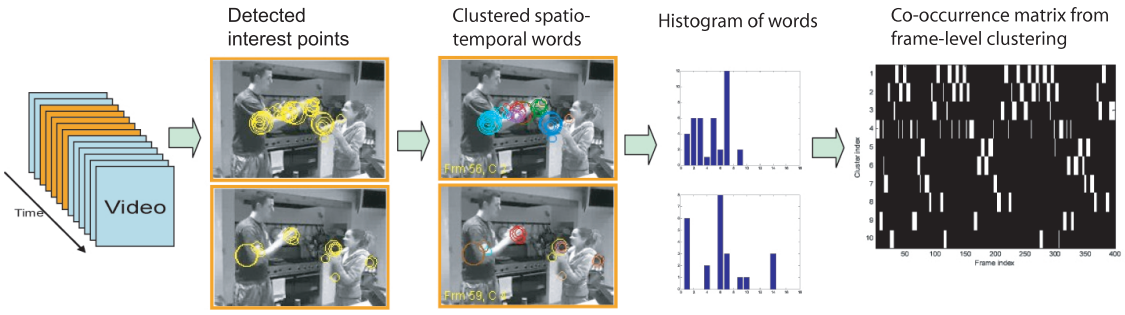



**Figure 4:** Illustration of our pipeline for converting an input video into a discrete symbolic sequence.

repetitions over a sequence of actions. Without the need to decompose the interactions in a social game into pre-defined actions, and the need for supervised learning, our algorithm can discover sequential actions that often correspond to the meaningful parsing of the game stages. This is a result of exploiting the temporal structure of social games.

Our approach consists of two stages. The first stage, illustrated in Figure 4, converts the input video into a discrete symbol sequence. The second stage analyzes this sequence and extracts quasi-periodic patterns. Within the first stage, tokenization of the video proceeds in two steps. First, extracted space-time features are clustered across the segment to form visual words. Second, each frame is represented by the histogram of its visual words, and the frame-level histograms are clustered to obtain keyframes. Then each frame is mapped to the nearest keyframe, yielding a sequence of keyframe indices. The second stage mines the quasi-periodic patterns from this sequence.

## *4.1 Discrete Sequence Representation of Videos*

**Video segments** The input video is segmented into overlapping clips of length $t\_win$ using a sliding window. Each segment overlaps 50% in number of frames with adjacent segments. $t\_win$ should be sufficiently long so that it covers at least two occurrences of the games. We used $t\_win = 500$ on average in our experiments. Each segment

(highlighted in orange in Figure 4) is then processed independently.

**Space-time interest points** Our discrete sequence representation is based upon space-time interest point detection proposed by Laptev [40]. Motivated by the success of sparse features in object recognition [47], sparse space-time features have shown promising performance in recognizing actions [43, 41, 42, 58, 37, 21, 75, 102, 94]. They are tolerant to unrestricted scenarios and large within-class variations due to difference in physical surroundings and human postures [43, 42].

The detector is an extension of Harris corner detection from 2D space domain to 3D spatio-temporal domain. Given a video sequence $v$, convolve it with a separable spatio-temporal Gaussian kernel $g(x, y, t; \sigma_l, \tau_l)$ to get $L = v * g(x, y, t; \sigma_l, \tau_l)$. The gaussian kernel is defined as

$$g(x, y, t; \sigma_l, \tau_l) = \frac{1}{(2\pi\sigma_l^2\sqrt{2\pi}\tau_l)} exp(-\frac{(x^2 + y^2)}{2\sigma_l^2} - \frac{t^2}{2\tau_l^2}) \tag{1}$$

At any point $p$, a second-moment matrix is defined as

$$\mu(p) = g(x, y, t; \sigma_i, \tau_i) * (\nabla L(p))(\nabla L(p))^T \tag{2}$$

$\nabla L = (L_x, L_y, L_t)^T$ is the gradient vector. $\sigma_i^2 = 2\sigma_l, \tau_i^2 = 2\tau_l$. Interest points are those have significant eigenvalues of $\mu$ ($\lambda_1, \lambda_2 or \lambda_3$). They are detected by finding local maxima of the generalized Harris function, defined as

$$H = det(\mu) - k\text{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \tag{3}$$

We set $k \approx 0.005$, following the choice of [40]. Similar to [42], we adopt a multi-scale approach to extract interest points at multiple spatio-temporal scales $\sigma_l = s_x\sigma_{l0}$, $\tau_l = s_t\tau_{l0}$, where $s_x^2 = 2^{i+1}, s_t^2 = 2^j$ with $i = 1, 2, 3, 4$ and $j = 1, 2$, $\sigma_{l0}^2 = 4, \tau_{l0}^2 = 2$. All the points with local maxima Harris scores are sorted in a descending order, then the top 30% of the sorted points are selected as the final detections $P$ (shown in yellow circles in Figure 4).

**Spatio-temporal words** Each interest point is represented by $(x, y, t, \sigma_l, \tau_l)$. Feature vector $f_p$ for each $p \in P$ has two components: a position-dependent histogram of oriented optical flows (Hof) from $p$'s space-time neighborhood and normalized $(x, y)$. It characterizes local motion and appearance, as well as the positional information. The size of the neighborhood $(\pm\Delta x, \pm\Delta y, \pm\Delta t)$ is determined by the detection scales $\sigma_l, \tau_l$ with $\Delta x = \Delta y = k_x\sigma_l, \Delta t = k_t\tau_l$. The neighborhood is divided into 8 (2 for each dimension) sub-volumes, and $L_2$ norm histograms are computed for all the sub-volumes with 18 bins. When calculating histogram for each sub-volume, a sphere 3D Gaussian is imposed to weight contributions from the voxels according to their distances to the center of the volume [47]. In our experiments, we set $k_x = 2, k_t = 2$. Orientation of each flow vector is weighted by its magnitude. Therefore, $f_p$ has 146 dimensions. Spatio-temporal words are built by applying K-means clustering to $\{f_p, p \in P\}$. The number of clusters is set to 16. Each interest point is assigned to a closest visual word (indicated by the colored circles in Figure 4).

**Frame descriptor** Each interest point is assigned to a closest visual word. Every frame is represented by the histogram of visual words in that frame. An interest point $p = (x, y, t, \sigma_l, \tau_l)$ contributes to the histograms over frame range $\{t-k_{ext}\tau_l : t+k_{ext}\tau_l\}$. We set $k_{ext} = 1.5$. We then assign each frame an event label by applying k-means clustering to the histogram representations of frames with $k = 10$. Each cluster center defines a keyframe, and corresponds to a histogram of visual words.

**Discrete sequence representation** The original video sequence $v$ is reduced to a sequence of event labels $s_e$. Co-occurrence matrix $M$ encodes the relationship between the video frames and the event labels. $M(i, j) = 1$ if $jth$ frame is assigned to event label $i$. We sort the events according to their cluster sizes in a descending order (cluster 1 has the biggest size). The $M$ shown in Figure 4 is for the patty-cake video in Figure 8. It is obvious that a subset of the events always occur close to each other and recur together in the same order.

Usually an event lasts several frames because the motion is continuous. To prepare for mining, sequence $s_e$ is compressed into $u_e$ as follows. Suppose the average number of continuous occurrences of event $e$ in the entire sequence is $n_e$, we only keep $e$ once if its current successive count $n > \max(1, n_e/5)$ and discard it otherwise. For example, a sequence of $(4, 4, 4, 6, 6, 6, 6, 3, 7, 7, 7, 4, 4)$ will be converted to $(4, 6, 7, 4)$. This is an empirical choice and the main goal is to accommodate the variations in motion speed, and to eliminate extraneous short actions (which typically doesn't last long).

## *4.2 Quasi-Periodic Pattern Extraction*

Intuitively, the quasi-periodic patterns should not only repeat in the sequence, but also correspond to "important" moments in the game, which often correspond to the distinctive stages of a game. Our observations show that the crucial moments, such as the sudden appearances or disappearances of a face in a peek-a-boo game, and the rapidly flying ball in a ball game, often generate visual words, therefore the frame labels that occur much less frequently than the ones that are caused by camera shaking, or other persistent background motions. The frame labels that are dominated by the crucial movements should carry more weight when characterizing the importance of the recurring patterns. Inspired by InfoMiner [98], we define pattern information $I(Pat)$ and a pattern score function $G(Pat)$ to measure the importance of a pattern. Each event $e$ is associated with certain information $I(e)$. We define $I(e) = -log_{|E|}p(e)$, where $|E|$ is the total number of events, and $p(e)$ is the frequency of $e$ in the sequence $s_e$. The more frequently an event occurs, the less information it carries. $I(Pat)$ is the sum of the *unique* events' information in $Pat$. This definition prevents a pattern consisting of repeated events (*e.g.*, $Pat = (1, 2, 1, 2, 1, 2)$) from having high information. The pattern score is defined as $G(Pat) = I(Pat) * (Occur(Pat) - 1)$, where $Occur(Pat)$ is the number of occurrences of $Pat$ in $u_e$. A pattern that appears only once has zero score. Note that these are heuristic choices designed to capture a

pattern's importance.

Algorithm 1 summarizes our quasi-periodic pattern extraction method. First, the recurring pattern set $Patset$ ($Occur(Pat) > 1$, $Pat \in Patset$) is extracted via a suffix tree [50]. Suffix tree is a data structure first proposed by Weiner [88] for organizing and comparing strings. We use the space-efficient method by McCreight [50] to build a suffix tree for each $u_e$ and traverse the tree to find the repeating patterns. Appendix A gives the intuitive explanation of how to build a suffix tree for a given string, and how to extract the repeating patterns from the suffix tree. A pattern is a path from the root to a node (which can be either an internal node or a leaf node) in the tree. The number of occurrences of that pattern is the number of leaf nodes of the subtree of that pattern. Suffix tree was used to detect abnormal activities [33]. Their sequences of activities are either generated by manual labeling of the events, or by event detection based on background subtraction from videos recorded with a fixed overhead camera (where the ground layout is known and the location of the moving target indicates a certain event). In contrast, our event sequence is generated automatically from very challenging videos, without any prior knowledge of the behaviors or the environment. Second, pattern information $I(Pat)$ and pattern score $G(Pat)$ are computed and compared against minimum score $min\_gain$ to decide if it is a valid quasi-periodic pattern.

**Algorithm 1** Quasi-periodic pattern extraction

Input: $u_e$, $min\_gain$
Output: Quasi-periodic patterns $QuasiPatSet$.

$(Patset, Occur(Pat)) \leftarrow$ `SuffixTree`$(u_e)$
**for** each $Pat = (e_1, e_2, \ldots, e_n) \in Patset$ **do**
  $ue \leftarrow$ `unique`$(e_1, e_2, \ldots, e_n)$ {$ue$ is the set of unique events in $(e_1, e_2, \ldots, e_n)$.}
  $I(Pat) \leftarrow \sum_i (I(ue(i)))$
  $G(Pat) \leftarrow I(Pat) * (Occur(Pat) - 1)$
  **if** $G(Pat) > min\_gain$ **then**
    $QuasiPatSet \leftarrow Pat$
  **end if**
**end for**

*QuasiPatSet* is a refinement of *Patset*. In the patty-cake example, suffix tree mining returns a *Patset* of 37 patterns, while the refined *QuasiPatSet* has only 11 patterns with $min\_gain = 1$. Among them, we highlight *Pat* 2-4-9-6, which occurred 3 times. As illustrated in Figure 8, the mined symbols map to stages of the game. More such examples will be shown in the experiments (Section 4.3). The results on parsing game stages and the game retrieval demonstrate the effectiveness of the pattern scoring function at picking characteristic patterns.

## *4.3 Experimental Results*

We present three different evaluations of our method for identifying quasi-periodic motion patterns. First, we examine the mined patterns from video clips of social games, demonstrating that the mined symbols correspond to meaningful game stages. Second, we demonstrate retrieval of social games from videos of children's play captured in the lab. Third, we explore the effectiveness of our method in retrieving general social interactions from home movies. The videos and associated visualizations from this section are available from our project website.

### 4.3.1 Mined patterns from videos of social games

Given a video of a social game, we demonstrate that our method can find patterns that correspond to meaningful stages of the games, for a variety of game types. Figure 7 shows the mined pattern and the correspondent frames for a peek-a-boo game. *Pat* 2-1-5-1-7 depicts the process from the toy fully appearing to fully disappearing and the associated responses from the baby. Event 2 is where the toy is held closest to the baby; event 5 is where the toy is half-hidden but still in the baby's view, and the baby is reaching for the toy; In event 7, the baby reaches out furthest for the hidden toy. Note that the duration for the two occurrences of the same pattern are quite different. It takes 25 frames from Figure 7(a) to Figure 7(e), and 72 frames from Figure 7(f) to Figure 7(j), since the father controls the climax of the baby's laughter

(a) Frame 465, label 2 (b) Frame 469, label 9 (c) Frame 473, label 7

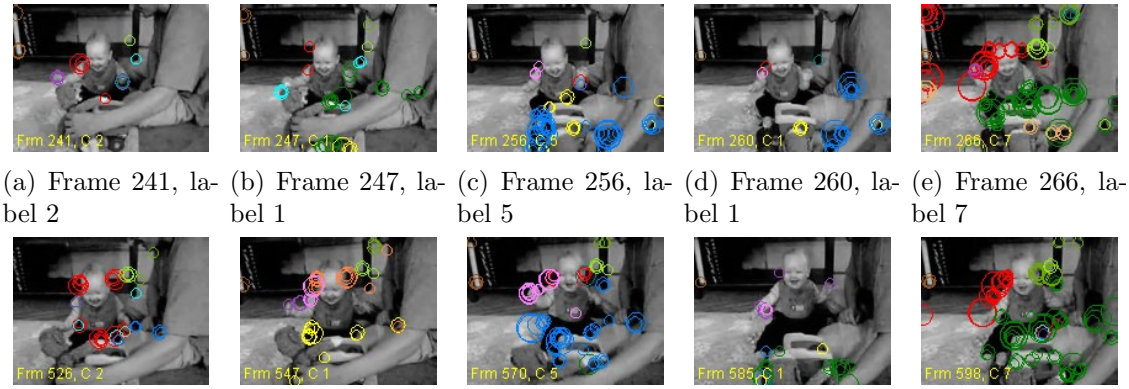
(d) Frame 653, label 2 (e) Frame 664, label 9 (f) Frame 668, label 7

**Figure 5:** Mined quasi-periodic pattern 2-9-7 and its two occurrences from YouTube video PeekabooMomDad. Interest points are color-coded to show their visual words assignments.

(a) Frame 144, label 4 (b) Frame 154, label 1 (c) Frame 171, label 2

(d) Frame 325, label 4 (e) Frame 334, label 1 (f) Frame 351, label 2

**Figure 6:** Mined quasi-periodic pattern 4-1-2 and its two occurrences from YouTube video PeekabooBabyDad. Label 4 is when dad is fully invisible to baby. Label 1 is dad shows his head upwards. Label 2 is dad hides himself in front of baby.




(a) Frame 241, label 2 (b) Frame 247, label 1 (c) Frame 256, label 5 (d) Frame 260, label 1 (e) Frame 266, label 7

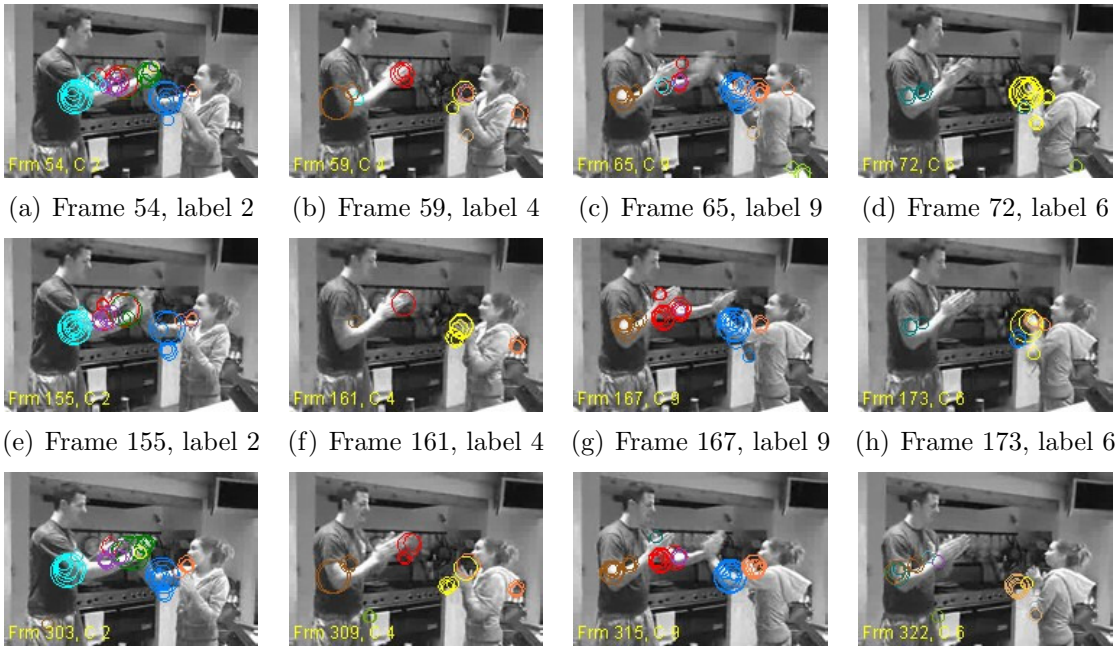
(f) Frame 526, label 2 (g) Frame 547, label 1 (h) Frame 570, label 5 (i) Frame 585, label 1 (j) Frame 598, label 7

**Figure 7:** Mined quasi-periodic pattern 2-1-5-1-7 and its two occurrences from YouTube video PeekabooMonkey. Label 2 - monkey toy appears. Label 1 - baby reaches for toy. Label 5 - toy is hidden. Label 7 - baby reaches further for toy.

by changing the rhythm of the game.

The example demonstrates empirically that the shortened sequence $u_e$ can handle the variations in motion speed. Readers may notice that interest points (in dark brown) occasionally appear in the background. These are caused by camera jitter (the videographer was holding the camcorder in his hand). Since the games are largely in-focus and the video contains only the game, those spurious points usually won't dominate the visual words in that frame.

Figure 8 shows the three occurrences of *Pat* 2-4-9-6 mined from a patty-cake video. *Pat* 2-4-9-6 depicts clapping right hands (label 2), withdrawing and clapping one's own hands (label 4), clapping left hands (label 9), withdrawing and clapping one's own hands again (label 6). This patty-cake game is played with a rapid and structured rhythm, so the durations of the occurrences are close to each other. The interval between the three occurrences are different. It takes 100 frames from Figure 8(a) to 8(e), and 148 frames from Figure 8(e) to 8(i). This is not surprising since the two players clap right/left hands once for the first turn, then twice for the second turn, and keep increasing the times of clapping as the turn increases.



(a) Frame 54, label 2 (b) Frame 59, label 4 (c) Frame 65, label 9 (d) Frame 72, label 6

(e) Frame 155, label 2 (f) Frame 161, label 4 (g) Frame 167, label 9 (h) Frame 173, label 6

(i) Frame 303, label 2 (j) Frame 309, label 4 (k) Frame 315, label 9 (l) Frame 322, label 6

**Figure 8:** Mined quasi-periodic pattern 2-4-9-6 and its three occurrences from YouTube video patty-cake. Label 2 - clap right hands. Label 4 - withdraw right hands and clap own hands. Label 9 - clap left hands. Label 6 - withdraw left hands and clap own hands.

Figure 9 further illustrates how the visual words contribute to *Pat* 2-4-9-6. There are four rows in the left image. Each row is a histogram of words representing a keyframe (cluster center), and is ordered as 2,4,9,6 from top to bottom. The horizontal axis is the word index. The brighter the histogram bin, the more frequently the visual word appears in the cluster. The right image shows the co-occurrence matrix $M_1$ between visual words and frames.[1] This illustrates the existence of quasi-periodic patterns for groups of visual words, such as the dominant words 5,6,7, and 14.

### 4.3.2 Social game retrieval experiment

For the experiment on social game retrieval, we recorded 3 sessions of parent-child free play with a hand-held camcorder in our child study lab. This dataset was designed

[1] $M_1(i,j)$ = sum of the occurrences of word i over frame range $\{j - k_{ext}\tau_l : j + k_{ext}\tau_l\}$. In contrast, the matrix $M$ gives the co-occurrence between temporal event labels (keyframes) and frames.

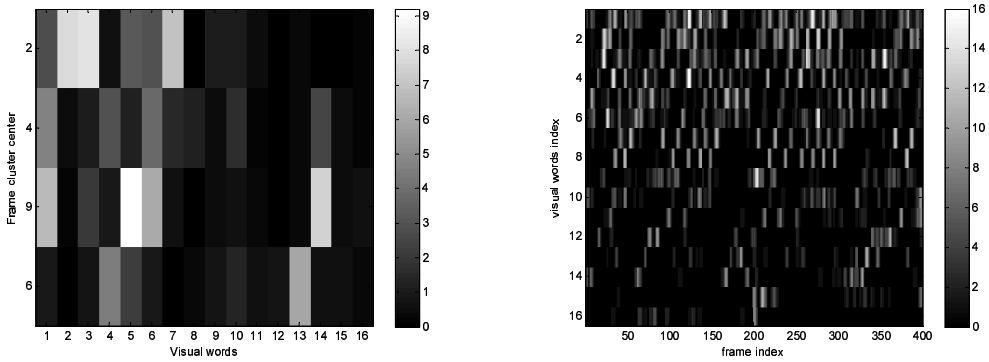



**Figure 9:** Histogram features for the pattern 2-4-9-6 from the patty-cake video. Left: cluster centers (keyframes) for 2,4,9,6. Right: co-occurrence matrix $M_1$ between visual words and frames.

to mix examples of social games with other kinds of play and non-play interactions. Thus it is a useful testing set for evaluating methods for social game retrieval. The video content is summarized in Table 2. This video dataset is a subset of our collection of children's play, which is called ChildPlay (Section 4.5). There are three videos corresponding to three adult-child pairs. The children were between 2 and 4 years old. The adults were instructed to freely mix social games with less structured interactions during the sessions. We asked them to include the games roll-the-ball, peek-a-boo and patty-cake. During the recording session, the videographer tried to keep the interactions centered within the camera view. The videos frequently display small amounts of camera motion. Not all the listed games were played by all the children, due to their own interests. Some new games were introduced during the sessions, resulting in the following additions to our final list of social games: "playing-drum-in-turn", "bowling", "give-me-five", "tickle", "give-and-take", "hot-potato", "frisbee" and "knock-hat-down". Ground truth labels were manually specified by the experimenter. A video segment was labeled as *game* only if a game interaction occurred at least twice in that segment, and was otherwise labeled *nongame*. Figure 10 illustrates some selected game instances from the three videos.

The retrieval performance was measured in terms of the precision-recall curve.

**Table 2:** Summary of videos for game retrieval experiment. Videos were segmented into 500-frame long windows and manually labeled.

| Video | Length (min) | #game | #nongame | $P_{manual}$ |
|---|---|---|---|---|
| 1 | 26 | 42 | 142 | 22.83% |
| 2 | 19 | 48 | 85 | 36.09% |
| 3 | 40 | 102 | 181 | 36.04% |

(a) drum (b) bowling (c) patty-cake (d) frisbee (e) peekaboo

**Figure 10:** Selected instances of social games in the three videos.

Each sliding window were ranked according to mean pattern score in the mined *QuasiPatSet* from that video segment. A (precision, recall) pair is calculated at each position in the ranked list.

Precision and recall are defined as follows:

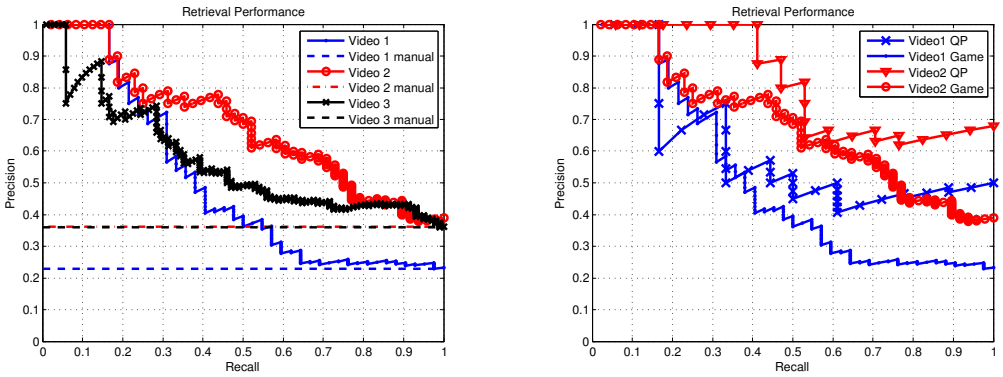
$$Precision = \frac{\#\text{true positives}}{\#\text{retrieved items}} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false positives}} \tag{4}$$

$$Recall = \frac{\#\text{true positives}}{\#\text{relevant items}} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false negatives}} \tag{5}$$

We analyzed the video for both the retrieval of social games and the retrieval of quasi-periodic events. The precision-recall curves are shown in Figure 11.

**Retrieval of Social Games**

In retrospective video study (Chapter 2, Section 2.2), it is often the case that after video editing, which means the entire video collection is viewed and relevant clips are marked, only a subset of the labeled video clips are randomly selected for detailed behavior coding in order to save time and labor. If we denote the precision for manual content editing as $P_{manual}$, then $P_{manual} = \#game/(\#game + \#nongame)$. If all the labeled clips are used for behavior coding, then $R_{manual} = 1$. In reality, $R_{manual} < 1$ due to subsampling, while $P_{manual}$ remains the same statistically (as




(a) Retrieval of Games

(b) Retrieval of QP events

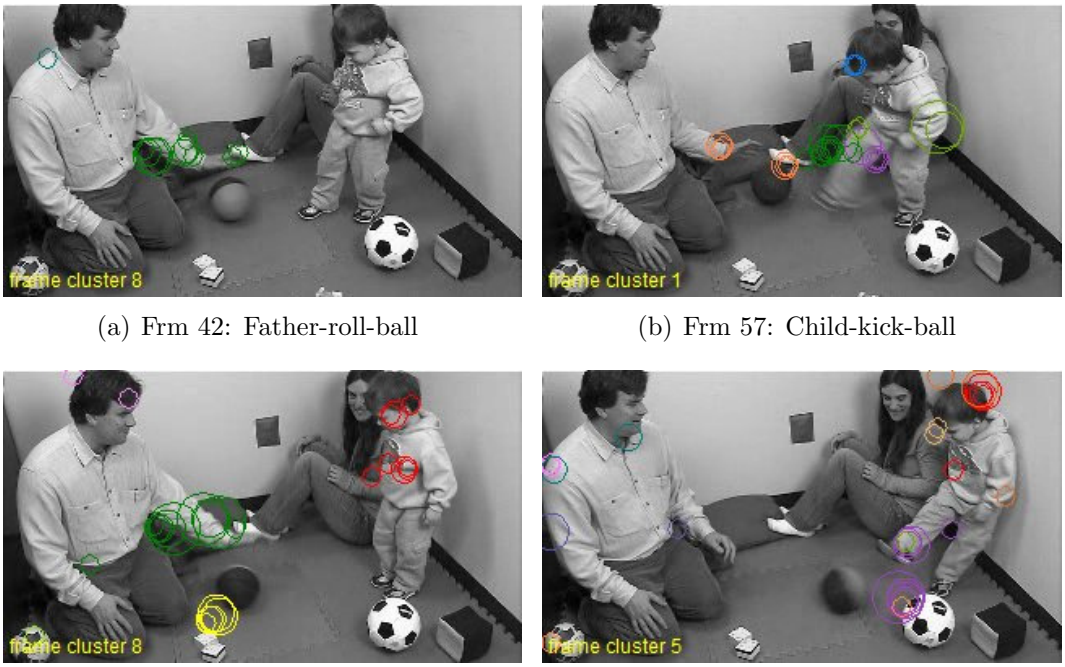
**Figure 11:** Precision (Y axis) - Recall (X axis) curves.

the content editing was conducted on the entire video collection). For the ChildPlay dataset, we have $P_{manual} = 22.83\%$, 36.09% and 36.04% respectively.

Figure 11(a) shows the precision-recall curves for the three videos. Our retrieval scheme yields a precision close to $P_{manual}$ when recall approaches 100%, but the precision doubles when the recall drops to 48% for video 1, 46% for video 2, and 28% for video 3. The results show promise for filtering video to efficiently identify examples of social games in unstructured collections.

**Retrieval of Quasi-Periodic Events**

Figure 11(b) gives a comparison of the retrieval performance for games and QP sequences on videos 1 and 2. A segment is labeled as QP if any human motion happens at least twice (with variations), such as a child banging toys on the floor. Two coders with no knowledge of this project independently labeled the same uniformly sampled subset of the original videos. Their interrater reliability is measured with Cohen's kappa coefficient, scoring 0.73 and 0.7 for videos 1 and 2, respectively, which represents substantial agreement. A third coder made a final decision on the segments with disagreements. Cohen's kappa measures interrater agreement between two raters for qualitative (categorical) items. ICC was used to measure interrater reliability for real-valued data such as frequency and rating scores in the retrospective study [7, 16].




(a) Frm 42: Father-roll-ball (b) Frm 57: Child-kick-ball

(c) Frm 275: Father-roll-ball (d) Frm 300: Child-kick-ball

**Figure 12:** A False Negative. The child kicked the ball with two very different poses, and the correspondent frames received label 1 and 5 respectively.

Cohen's kappa coefficient $\kappa$ is defined as

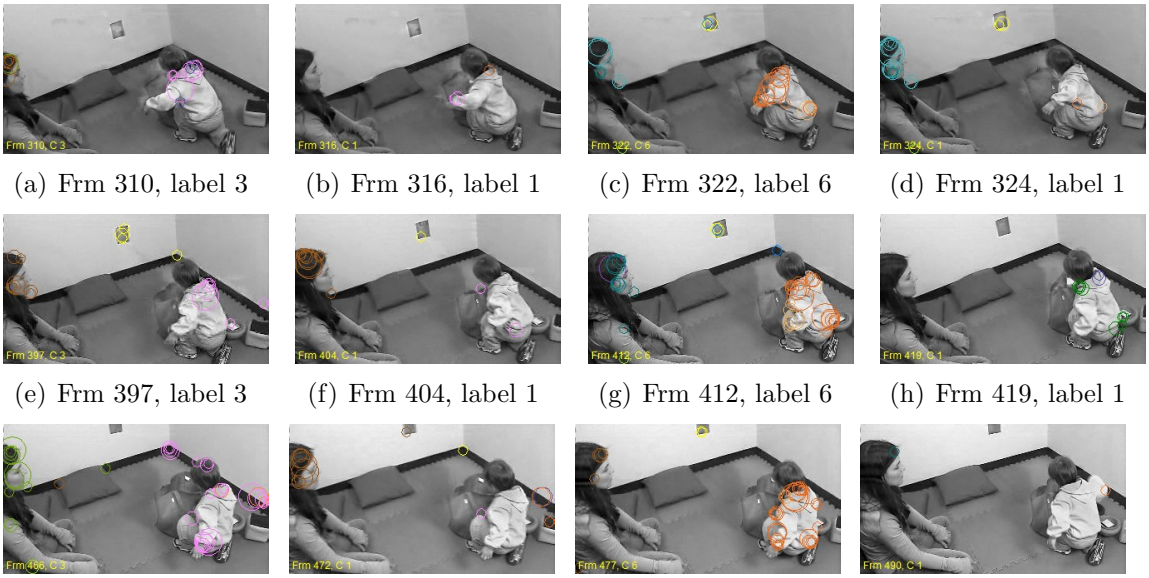
$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{6}$$

where $Pr(a)$ is the relative observed agreement, and $Pr(e)$ is the hypothetical probability of chance agreement.

For video 1, at 50% recall, the precision is 54% for QP, and it is 40% for game; at 100% recall, the precisions are 50% and 23% for QP and game respectively. For video 2, at recall 100%, the precision is 68% for QP, and 39% for game. The better retrieval performance for QP is not surprising since our algorithm is designed to detect QP events and hasn't addressed the turn-taking interactions in games.

**Analysis of True Positives (TP), False Positives (FP) and False Negatives (FN)**

We show some of the TP, FP and FN examples from the game retrieval experiment.



(a) Frm 310, label 3 (b) Frm 316, label 1 (c) Frm 322, label 6 (d) Frm 324, label 1

(e) Frm 397, label 3 (f) Frm 404, label 1 (g) Frm 412, label 6 (h) Frm 419, label 1

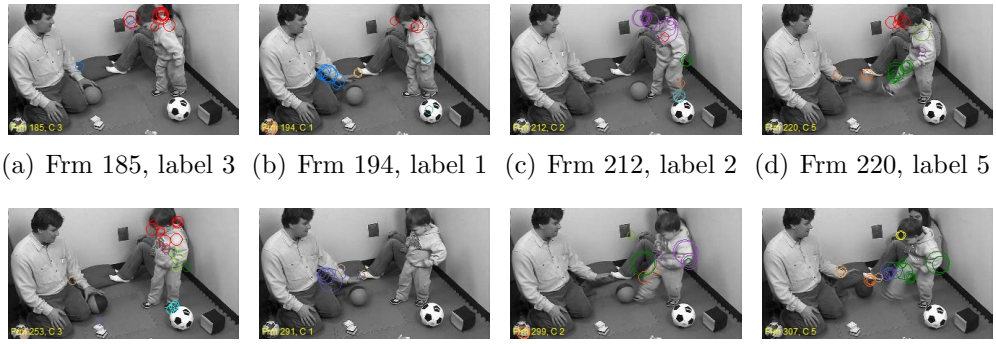
(i) Frm 466, label 3 (j) Frm 472, label 1 (k) Frm 477, label 6 (l) Frm 490, label 1

**Figure 13:** A False Positive. It is a sequence of repetitive actions without interacting with the other person in the scene.

Figure 12 shows a FN where two occurrences of father-roll-the-ball followed by child-kick-the-ball are missed. The main cause is that the child kicked the ball with two very different poses, and as a result these frames received different labels (1 and 5), in spite of the fact that the two roll-the-ball actions from the father were assigned the same cluster label 8.

Figure 13 gives an example of a FP. The child is taking out toys from a bag one by one. It is indeed a repetitive sequence and our algorithm finds pattern 3-1-6-1 occurs three times in the video.

Figure 14 shows two occurrences of pattern 3-1-2-5 of a TP example. Event 3 is dad-prepare-to-roll. Event 1 is roll-ball. Event 2 is child-prepare-to-kick. Event 5 is child-kick. Both the kicking poses and the distance between the dyad are quite similar in the two occurrences, so our algorithm can assign labels correctly.



(a) Frm 185, label 3 (b) Frm 194, label 1 (c) Frm 212, label 2 (d) Frm 220, label 5

(e) Frm 253, label 3 (f) Frm 291, label 1 (g) Frm 299, label 2 (h) Frm 307, label 5

**Figure 14:** A True Positive.

### 4.3.3 Home movie experiment

In this experiment, we evaluate our algorithm to retrieve more general social interactions from our home movie collection. We chose the first 9 years (which were recorded during 1957-1965) in the archived home movies as our testing videos, as it features a lot of child-child and adult-child interactions. This video corpus is a very challenging dataset for computer vision analysis. It is significantly more complex than the vast majority of the datasets currently in use for activity recognition. Our choice of $t\_win = 300$ produces 2240 segments, approximately 3.75 hours with a frame rate of 25 fps. Typical scenarios in the archive include Christmas parties, birthdays, summer vacations, at church, outdoor playing and school activities. Out of the entire footage, there are only 5 segments containing ball games (with 50% overlap), and there is no occurrence of peek-a-boo and patty-cake. On the other hand, the archive does contain many social interactions between the children and the adults. In this experiment, we retrieve recurring social interactions.

A video segment is labeled as *interaction* when there are recurring social interactions between a child and an adult, or among any number of children. Two conditions must be met for an *interaction* label: 1) both players (or the body parts participating in the interactions) are in the view; 2) the interaction must occur at least twice, with

actions that are similar. Therefore, two-person conversation is labeled as *interaction* if an actor makes gestures repetitively, and is labeled as *noninteraction* if both actors remain static. Two boys helping each other to unwrap a large present is an *interaction* (with many reach-and-tear actions), while later examples of them opening their own gifts are not, because they don't interact. This criteria results in 261 segments of *interaction* and 1979 segments of *noninteraction*.

The retrieval performance is shown in Figure 15. The red dotted horizontal line is the precision-recall curve of manual labelling ($P_{manual} = 11.65\%$ for any recall). Our precision is higher than manual labelling when the recall rate is between 40% and 90%, and much higher when recall is below 7%. Overall there is significant room to improve the retrieval performance on this challenging dataset.

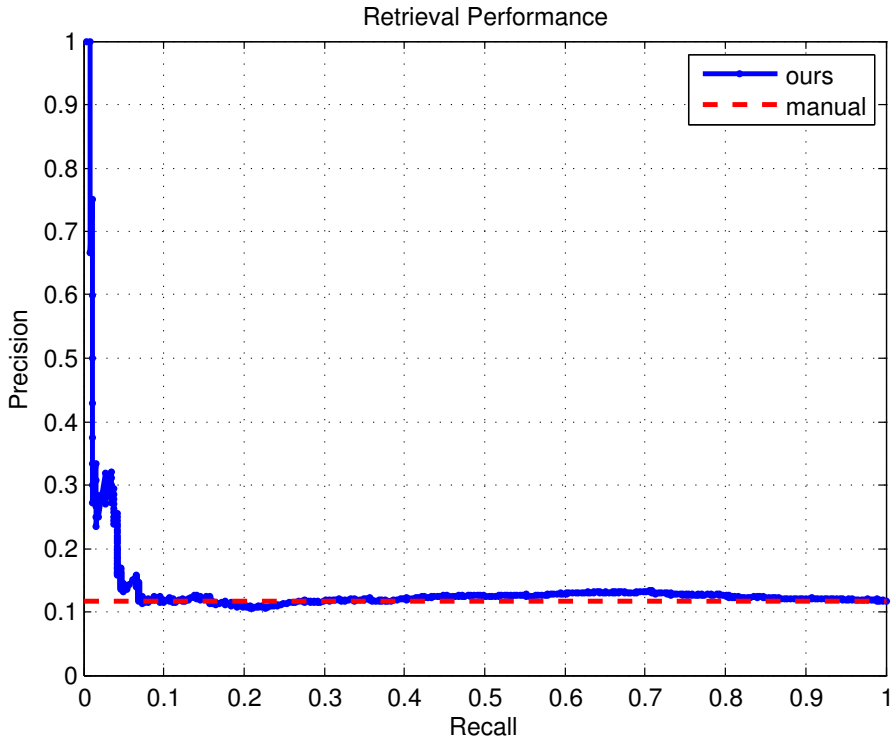



**Figure 15:** Retrieval performance.

Examples of TPs and FPs are shown in Figure 16. There are mainly 4 types of FPs: 1) children doing similar actions one by one. For example, hanging stockings

(a) TP: playing doll

(b) TP: wrestling

(c) TP: playing with baby

(d) TP: badminton

(e) TP: toss ball

(f) FP: driving in a go-cart

(g) FP: dinner

(h) FP: hanging socks

(i) FP: taking exam

(j) FP: praying

(k) FP: posing

(l) FP: snow

(m) FP: outdoors

(n) FP: entering door

(o) FP: at church

(p) FP: crowd

**Figure 16:** Selected true positives (TP) and false positives (FP).

at Christmas (Figure 16(h)), boys being inspected before swimming (Figure 16(i)). In these examples, the camera view is very similar for each child. The body inspection example is an interaction, but it is not a repetitive interaction from the same dyad. 2) A group of people posing or sitting together and making small movements (Figure 16(j), 16(k)). 3) parallel play, where two people are in view at the same time, but do not interact (Figure 16(l)). 4) static scenes recorded by a (smoothly) moving camera (Figure 16(m)). We believe that further analysis of the turn-taking interactions should reduce the FPs.

## *4.4 Parameter Sensitivity Study*

We have shown the efficacy of our two-stage unsupervised approach at extracting quasi-periodic patterns that correspond to social games from unstructured videos. Once the frame labels are assigned to all the frames in the sliding window, the quasi-periodic patterns are mined deterministically. As a result, the frame label assignment has an impact on the pattern mining. The goal of the frame label assignment is to assign same labels to the frames that contain similar actions, which are represented by histograms of the visual words in that frame (and from its adjacent frames). If the total number of frame labels $k_{event}$ is too big, similar actions will be assigned with different labels, and the QP pattern mining may not find recurring patterns; if $k_{event}$ is too small, one label may be assigned to frames of different actions, and the QP pattern mining may find short patterns that repeat often. The histogram of visual words at any given frame depends on the number of visual words $k_{word}$, assuming that the features of interest points $\{f_p, p \in P\}$ characterize the local motion and appearance around each interest point $p$ well. $k_{word}$ determines how the interest points are grouped. We expect the interest points that are generated by the same action should be grouped into one visual word. If $k_{word}$ is too small, interest points generated by different actions will be grouped into one word; if $k_{word}$ is too big,

similar actions may generate interest points that are clustered into different words. Both cases will affect the frame label assignment.

From the analysis above, we conclude that the choice of $k_{event}$ and $k_{word}$ are important to the QP pattern mining, therefore the retrieval performance. In this section, we will empirically evaluate the quasi-periodic pattern mining and retrieval performance on a variety choices of $k_{word}$ and $k_{event}$. Our baseline choices for these two parameters are $k_{word} = 16$ and $k_{event} = 10$, which were used in all the previous experiments. We then fix either one of them, and increase/decrease the other parameter up to 20% of the baseline value, resulting in the parameter values shown in Table 3. It is hypothesized that the variations of $k_{event}$ will have bigger impact on the retrieval performance than that of $k_{word}$, and that bigger $k_{event}$ tends to give more detailed parsing of social game stages.

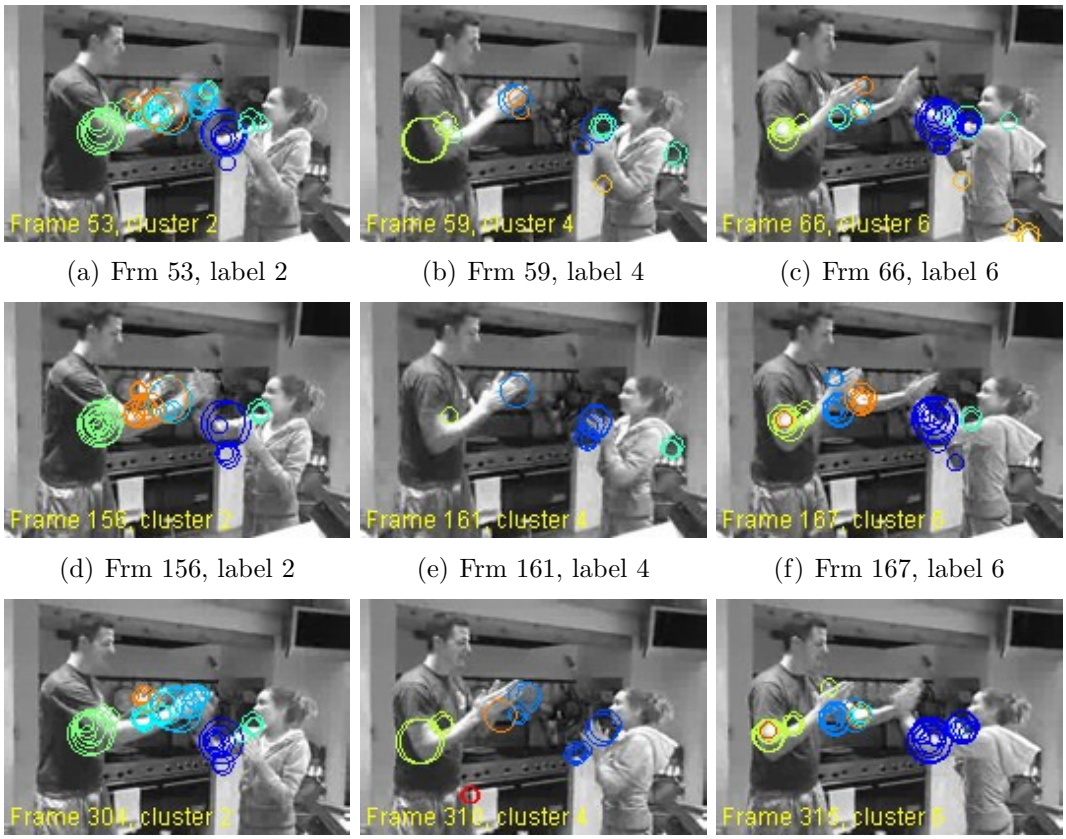
**Table 3:** Values of $k_{word}$ and $k_{event}$ in parameter sensitivity test.

| | |
|---|---|
| $k_{event} = 10$ | $k_{word} = (13, 14, 15, 16, 17, 18, 19)$ |
| $k_{word} = 16$ | $k_{event} = (8, 9, 10, 11, 12)$ |

### 4.4.1 Mine patterns from YouTube videos of social games

In this section, we evaluate the sensitivity of $k_{word}$ and $k_{event}$ at parsing the social games stages. We tested the parameter sensitivity on the YouTube videos of patty-cake and peek-a-boo.

First we evaluate the effect of varying $k_{word}$ when $k_{event}$ is fixed. We set $k_{event} = 10$. For the patty-cake video, Figure 17 shows the pattern 2-4-6 with $k_{word} = 13$, and Figure 18 shows the pattern 1-4-3-9-10 with $k_{word} = 19$. For the PeekabooBabyDad video, in which the father plays peekaboo with the baby by hiding and reappearing his face in front of the baby repetitively, the mined patterns 5-2-1-3 and 2-1-3 are shown in Figure 21 with $k_{word} = 13$ and Figure 22 with $k_{word} = 18$. All the patterns correspond to meaningful stages of the game. The parsing of game stages is not




(a) Frm 53, label 2 (b) Frm 59, label 4 (c) Frm 66, label 6

(d) Frm 156, label 2 (e) Frm 161, label 4 (f) Frm 167, label 6

(g) Frm 304, label 2 (h) Frm 310, label 4 (i) Frm 315, label 6

**Figure 17:** All the three occurrences of Pattern 2-4-6 mined from the YouTube patty-cake video with $k_{word} = 13, k_{event} = 10$. Label 2: clap right hands. Label 4: withdraw right hands. Label 6: clap left hands.

(a) Frm 48, label 1 (b) Frm 54, label 4 (c) Frm 60, label 3 (d) Frm 66, label 9 (e) Frm 73, label 10

(f) Frm 150, label 1 (g) Frm 155, label 4 (h) Frm 161, label 3 (i) Frm 167, label 9 (j) Frm 173, label 10

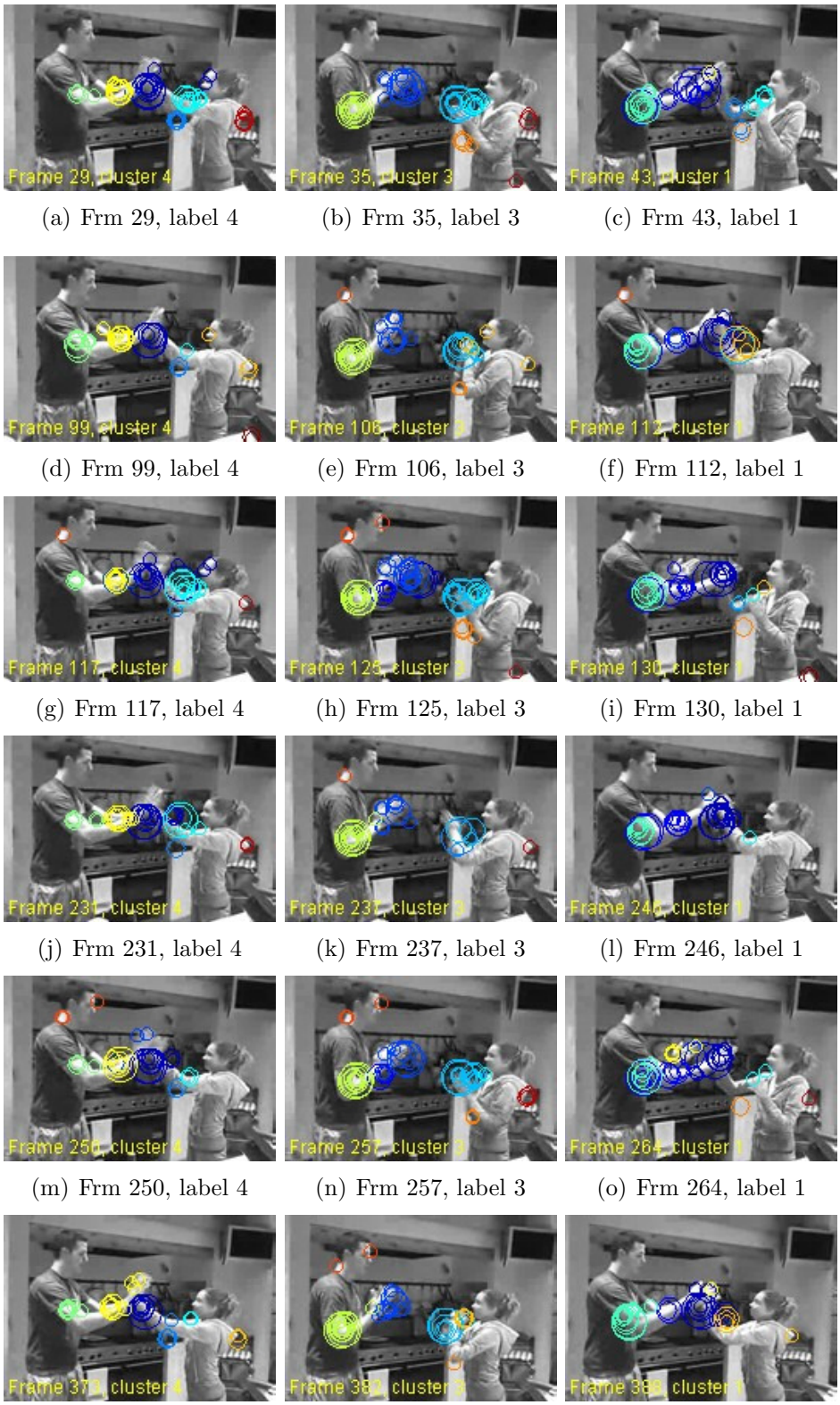
(k) Frm 296, label 1 (l) Frm 302, label 4 (m) Frm 309, label 3 (n) Frm 316, label 9 (o) Frm 322, label 10

**Figure 18:** All the three occurrences of Pattern 1-4-3-9-10 mined from the YouTube patty-cake video with $k_{word} = 19, k_{event} = 10$. Label 1: prepare to clap hands. Label 4: clap right hands. Label 3: withdraw right hands. Label 9: clap left hands. Label 10: withdraw left hands.

sensitive to the variations of $k_{word}$ once $k_{event}$ is chosen appropriately.

Next, we set $k_{word} = 16$, and varied $k_{event}$ according to Table 3. At the 20% range of variations around baseline $k_{event} = 10$, our algorithm was still able to get meaningful parsing of the game stages. Such parsing often captures more details of the game stages when $k_{event}$ is bigger. Figure 20 shows the pattern 5-7-3-2-1 with parameters $k_{word} = 16, k_{event} = 12$. Label 5 is clapping left hands; label 7 corresponds to withdraw and prepare for the next turn; label 3 is clapping both hands; label 2 is again clapping both hands; label 1 is withdrawing hands. Figure 24 shows the pattern 1-5-2-1-3 mined from the PeekabooBabyDad video. It captures the complete cycle of a peekaboo game from the moment when the baby is looking elsewhere (label 1), to father appears (label 5), and attracts the baby's attention (label 2), to the mutual gaze exchange (label 1), and the disappearance of the father (label 3).

Figure 19 shows the mined pattern 4-3-1 from the patty-cake video when $k_{word} =$




(a) Frm 29, label 4 (b) Frm 35, label 3 (c) Frm 43, label 1

(d) Frm 99, label 4 (e) Frm 106, label 3 (f) Frm 112, label 1

(g) Frm 117, label 4 (h) Frm 125, label 3 (i) Frm 130, label 1

(j) Frm 231, label 4 (k) Frm 237, label 3 (l) Frm 246, label 1

(m) Frm 250, label 4 (n) Frm 257, label 3 (o) Frm 264, label 1

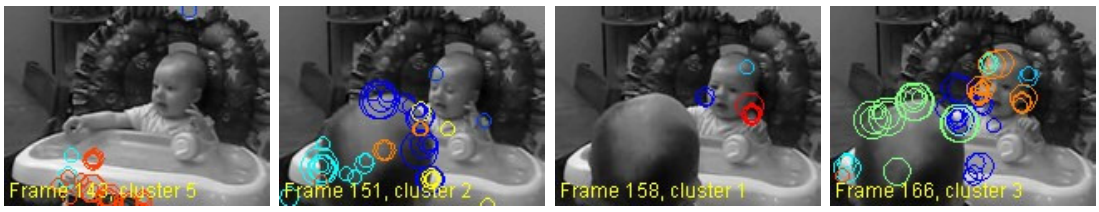
(p) Frm 373, label 4 (q) Frm 382, label 3 (r) Frm 388, label 1

**Figure 19:** All the six occurrences of Pattern 4-3-1 mined from the YouTube patty-cake video with $k_{word} = 16, k_{event} = 8$. See text for details.

(a) Frm 79, label 5 (b) Frm 87, label 7 (c) Frm 93, label 3 (d) Frm 97, label 2 (e) Frm 106, label 1

(f) Frm 354, label 5 (g) Frm 363, label 7 (h) Frm 368, label 3 (i) Frm 374, label 2 (j) Frm 382, label 1

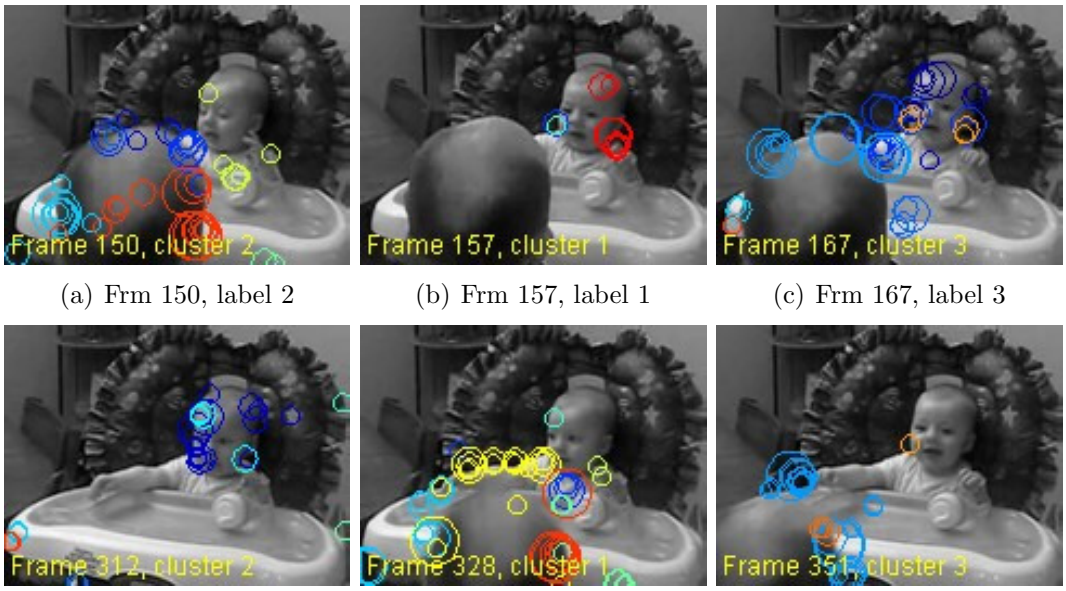
**Figure 20:** The two occurrences of Pattern 5-7-3-2-1 mined from the YouTube patty-cake video with $k_{word} = 16, k_{event} = 12$. See text for details.




(a) Frm 143, label 5 (b) Frm 151, label 2 (c) Frm 158, label 1 (d) Frm 166, label 3

(e) Frm 323, label 5 (f) Frm 331, label 2 (g) Frm 337, label 1 (h) Frm 343, label 3

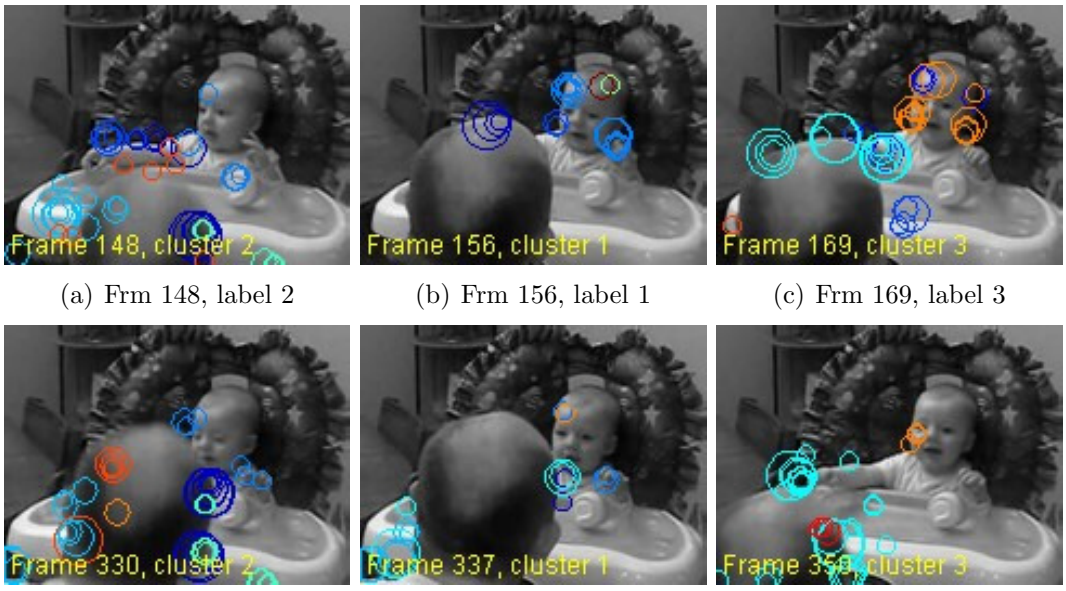
**Figure 21:** The two occurrences of Pattern 5-2-1-3 mined from the YouTube PeekabooBabyDad video with $k_{word} = 13, k_{event} = 10$. Label 5: baby looks at elsewhere. Label 2: father appears. Label 1: baby looks at father. Label 3: father disappears.




(a) Frm 150, label 2 (b) Frm 157, label 1 (c) Frm 167, label 3

(d) Frm 312, label 2 (e) Frm 328, label 1 (f) Frm 351, label 3

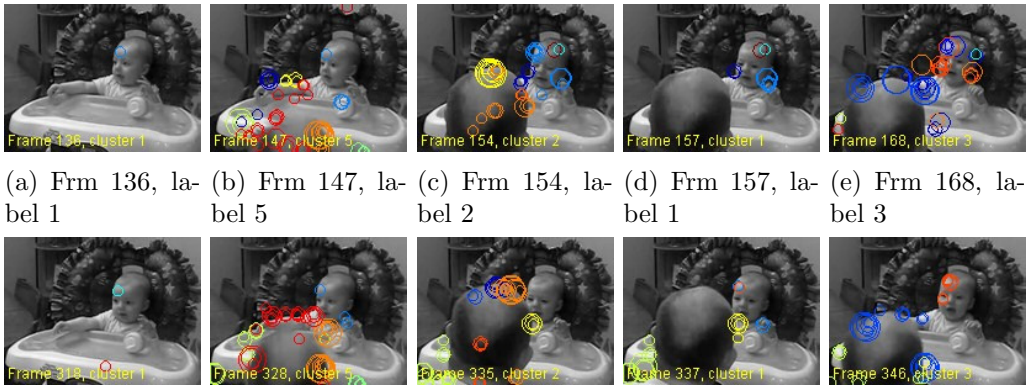
**Figure 22:** The two occurrences of Pat 2-1-3 mined from the YouTube video PeekabooBabyDad with $k_{word} = 18, k_{event} = 10$. Label 2: baby looks at elsewhere while father appears. Label 1: baby looks at father. Label 3: father disappears.




(a) Frm 148, label 2 (b) Frm 156, label 1 (c) Frm 169, label 3

(d) Frm 330, label 2 (e) Frm 337, label 1 (f) Frm 350, label 3

**Figure 23:** The two occurrences of Pat 2-1-3 mined from the YouTube video PeekabooBabyDad with $k_{word} = 16, k_{event} = 8$. Label 2: baby looks at elsewhere when father appears. Label 1: baby looks at father. Label 3: father disappears.




(a) Frm 136, label 1 (b) Frm 147, label 5 (c) Frm 154, label 2 (d) Frm 157, label 1 (e) Frm 168, label 3

(f) Frm 318, label 1 (g) Frm 328, label 5 (h) Frm 335, label 2 (i) Frm 337, label 1 (j) Frm 346, label 3

**Figure 24:** The two occurrences of Pat 1-5-2-1-3 mined from the YouTube video PeekabooBabyDad with $k_{word} = 16, k_{event} = 12$. Label 1: baby looks at elsewhere. Label 5: father appears. Label 2: baby turns his head/eyes towards father. 1: baby looks at father. 3: father disappears.

$16, k_{event} = 8$. Among all the six occurrences, label 4 depicts clapping both hands, and label 3 depicts withdrawing both hands together. Notice that label 1 corresponds to two different actions. In the 1st, 3rd and 5th occurrence, label 1 is clapping right hands. In the 2nd, 4th and 6th occurrence, label 1 is clapping both hands. Due to self-occlusions caused by the camera view, clapping both hands and clapping right hands generate similar visual word distributions. Given that the total number of frame labels is only 8 (probably less than sufficient), it is likely to assign the same frame label to the two actions. Figure 23 shows the pattern 2-1-3 from the PeekabooBabyDad video with $k_{word} = 16, k_{event} = 8$. It highlights the three stages of the game: baby looking at elsewhere (label 2), baby looking at the father when the father fully appears in front of the baby (label 1) and the father hides himself (label 3).

From the parameter test on the YouTube videos, we make the following conclusions:

1. The quasi-periodic mining algorithm is robust with respect to variations in $k_{word}$ and $k_{event}$. In particular, the variations of $k_{word}$ does not affect the meaningful parsing of the games once $k_{event}$ is selected properly.

2. When $k_{event}$ is small, we observe that sometimes two slightly different actions would be assigned to the same frame label (Figure 19). On the other hand, when $k_{event}$ is big, the mined patterns turn to give more detailed parsing of game stages (Figure 20 and 24). Overall, the mined patterns still often match to the meaningful stages of the game.

### 4.4.2 Social game retrieval

We evaluate the performance of social game retrieval on the ChildPlay dataset under all the choices of $k_{event}$ and $k_{word}$ listed in Table 3. The retrieval performance is measured in terms of the precision-recall curves, and the average precision (AP) (the area under the Precision-Recall curve).

When generating the precision-recall curve, each retrieved video segment was ranked according to three criteria: the mean pattern score $mean(G)$, the max pattern score $max(G)$, and the median pattern score $median(G)$ of the mined QP pattern set $QuasiPatSet$, where $\{G = G(Pat)|Pat \in QuasiPatSet\}$. If $QuasiPatSet$ is empty for a video segment, then the score is zero. For each pair of $k_{event}$ and $k_{word}$, we obtained 3 precision-recall curves.

Average precision measures precision, recall and the relevance ranking. The most relevant item has the highest pattern score. AP is the average of precision calculated at each position in the ranking list.

$$AP = \frac{\sum_{r=1}^{N} Precision(r) \times rel(r)}{\#\text{relevant items}} \tag{7}$$

where $r$ is the $rth$ hit in the ranking list, $N$ is the total number of retrieved items, $rel(r) = 1$ if $rth$ hit is relevant and 0 otherwise. $Precision(r)$ is the cutoff precision at rank $r$.

**Parameter Sensitivity Test on Video 1**

The precision-recall curves for Video 1 in ChildPlay dataset are shown in Figure 25, 26 and 27. Table 4 and 5 summarize the APs for each parameter pair that

**Table 4:** APs for Video 1 in ChildPlay dataset with varying $k_{event}$. $k_{word} = 16$. The highest AP is highlighted in bold.

| $k_{event}$ | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| $AP_{mean}$ | 0.3287 | 0.3022 | **0.3807** | 0.3651 | 0.3789 |
| $AP_{max}$ | 0.3239 | 0.3646 | **0.5049** | 0.3874 | 0.402 |
| $AP_{median}$ | 0.2936 | 0.2851 | 0.3027 | **0.3317** | 0.3157 |

**Table 5:** APs for Video 1 in ChildPlay dataset with varying $k_{word}$. $k_{event} = 10$. The highest AP is highlighted in bold.

| $k_{word}$ | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|
| $AP_{mean}$ | 0.380682 | 0.2682 | 0.302 | **0.380685** | 0.3749 | 0.2776 | 0.3316 |
| $AP_{max}$ | 0.3903 | 0.3349 | 0.298 | **0.5049** | 0.3103 | 0.3126 | 0.3504 |
| $AP_{median}$ | 0.3318 | 0.2481 | 0.3048 | 0.3027 | **0.3563** | 0.2618 | 0.3363 |

**Table 6:** APs for Video 2 in ChildPlay dataset with varying $k_{event}$. $k_{word} = 16$. The highest AP is highlighted in bold.

| $k_{event}$ | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| $AP_{mean}$ | 0.5663 | 0.6092 | **0.6839** | 0.6077 | 0.6472 |
| $AP_{max}$ | 0.5915 | 0.5208 | 0.642 | 0.642 | **0.6562** |
| $AP_{median}$ | 0.5116 | **0.6373** | 0.6214 | 0.5951 | 0.6357 |

**Table 7:** APs for Video 2 in ChildPlay dataset with varying $k_{word}$. $k_{event} = 10$. The highest AP is highlighted in bold.

| $k_{word}$ | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|
| $AP_{mean}$ | 0.5699 | 0.6152 | 0.5634 | **0.6839** | 0.5224 | 0.6116 | 0.5563 |
| $AP_{max}$ | 0.53 | 0.5581 | 0.5515 | **0.642** | 0.5023 | 0.5359 | 0.5757 |
| $AP_{median}$ | 0.5491 | 0.5957 | 0.5299 | **0.6214** | 0.505 | 0.5554 | 0.5487 |

**Table 8:** APs for Video 3 in ChildPlay dataset with varying $k_{event}$. $k_{word} = 16$. The highest AP is highlighted in bold.

| $k_{event}$ | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| $AP_{mean}$ | 0.5639 | **0.5921** | 0.5501 | 0.519 | 0.5122 |
| $AP_{max}$ | 0.5532 | **0.582** | 0.5724 | 0.5279 | 0.5204 |
| $AP_{median}$ | 0.5046 | **0.5137** | 0.4834 | 0.4945 | 0.488 |

were ranked by $mean(G)$, $max(G)$ and $median(G)$ respectively.

For Video 1, we observed that the pair $k_{event} = 10, k_{word} = 16$ gave the best AP (0.5049) when the retrieved items were ranked according to $max(G)$. The best parameter pair is $k_{event} = 10, k_{word} = 16$ for both $max(G)$ and $mean(G)$ based ranking.

**Table 9:** APs for Video 3 in ChildPlay dataset with varying $k_{word}$. $k_{event} = 10$. The highest AP is highlighted in bold.

| $k_{word}$ | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|
| $AP_{mean}$ | 0.5415 | 0.5711 | 0.5201 | 0.5501 | 0.5335 | **0.5916** | 0.5358 |
| $AP_{max}$ | 0.5402 | 0.5593 | 0.479 | 0.5724 | 0.5134 | **0.5792** | 0.5377 |
| $AP_{median}$ | 0.5105 | **0.5481** | 0.515 | 0.4834 | 0.5328 | 0.5262 | 0.5254 |

**Parameter Sensitivity Test on Video 2**

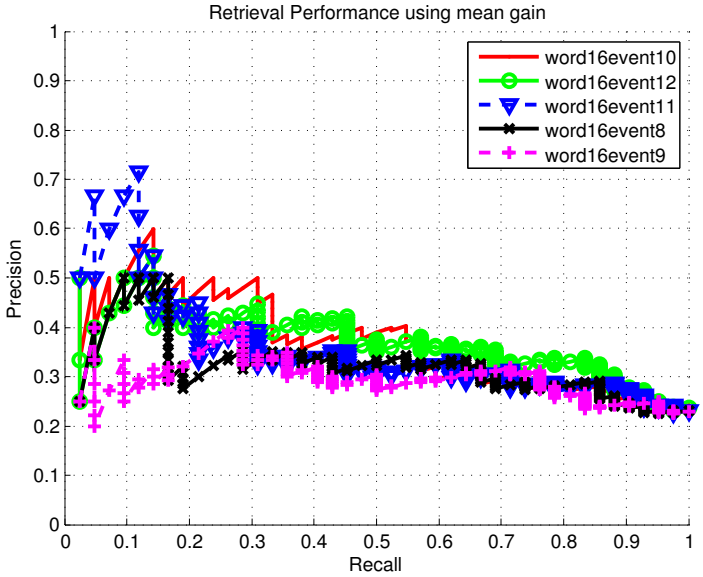
The precision-recall curves for Video 2 in ChildPlay dataset are shown in Figure 28, 29 and 30. Table 6 and 7 summarize the APs for each parameter pair that were ranked by $mean(G)$, $max(G)$ and $median(G)$ respectively.

We make two observations. First, $k_{event} = 10, k_{word} = 16$ gave the best AP (0.6839) when the retrieved items were ranked according to $mean(G)$. Second, the retrieval precision varies the most when recall is between 0 and 30%, and becomes close to each other when the recall is above 70%. In the practice of retrospective video study, where false positives are preferred over false negatives in the video content editing step, since behavioral scientists need to acquire as many as possible the relevant behaviors from the home movies, the retrieval performance is robust to the changes of $k_{event}$ and $k_{word}$.

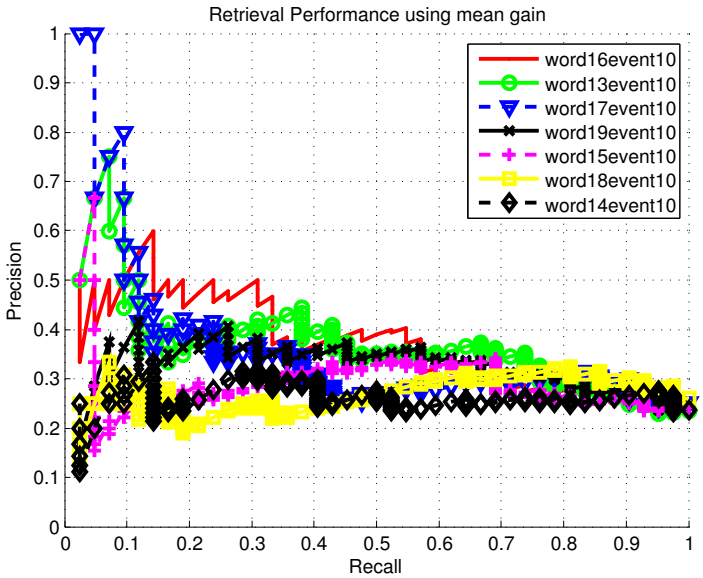
**Parameter Sensitivity Test on Video 3**

The precision-recall curves for Video 3 in ChildPlay dataset are shown in Figure 31, 32 and 33. Table 8 and 9 summarize the APs for each parameter pair that were ranked by $mean(G)$, $max(G)$ and $median(G)$ respectively.

In comparison to Video 1 and 2, variations of $k_{event}$ and $k_{word}$ make less impact on the retrieval performance. All APs and the precision-recall curves are close to each other. $k_{event} = 9, k_{word} = 16$ gave the best AP (0.5921) when the rank list was sorted according to $mean(G)$. We also observe that the rank lists sorted according to $max(G)$ and $mean(G)$ have better retrieval quality that the list sorted according to $median(G)$.




(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$

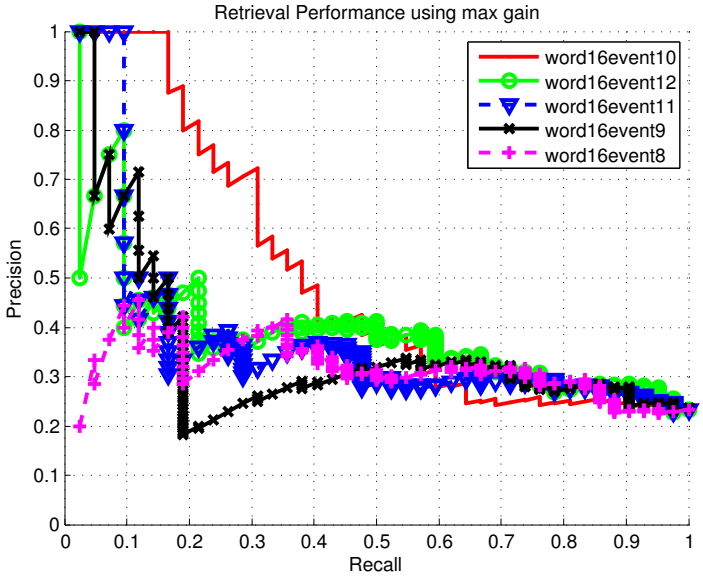



(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

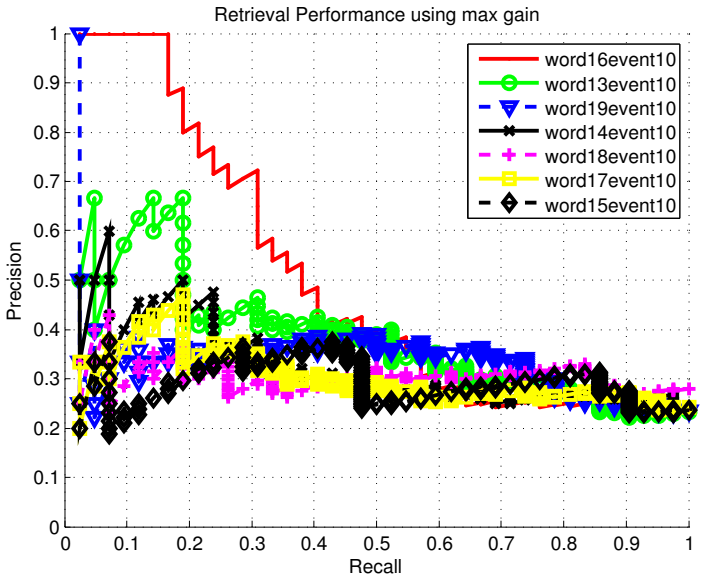
**Figure 25:** Precision-recall curve based on $mean(G)$ for Video 1 in ChildPlay dataset.

### Summary of parameter sensitivity study on social game retrieval

From the parameter sensitivity study on the ChildPlay dataset with the three ranking criteria, we make the following conclusions:




(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$

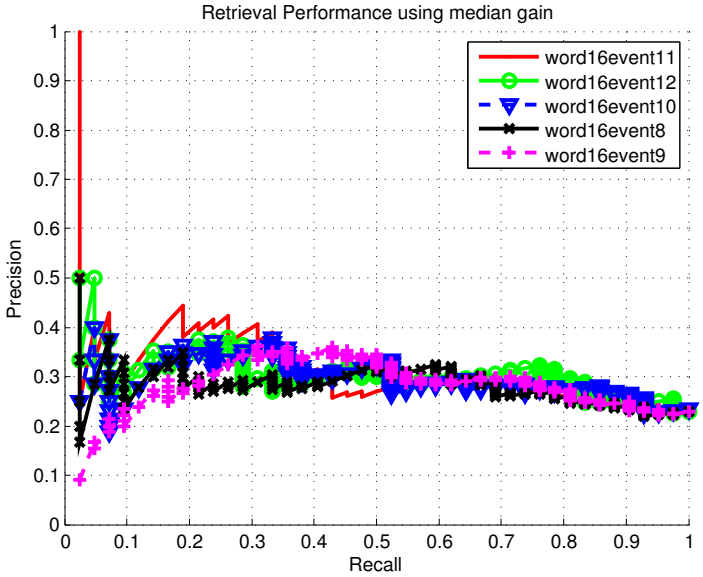



(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

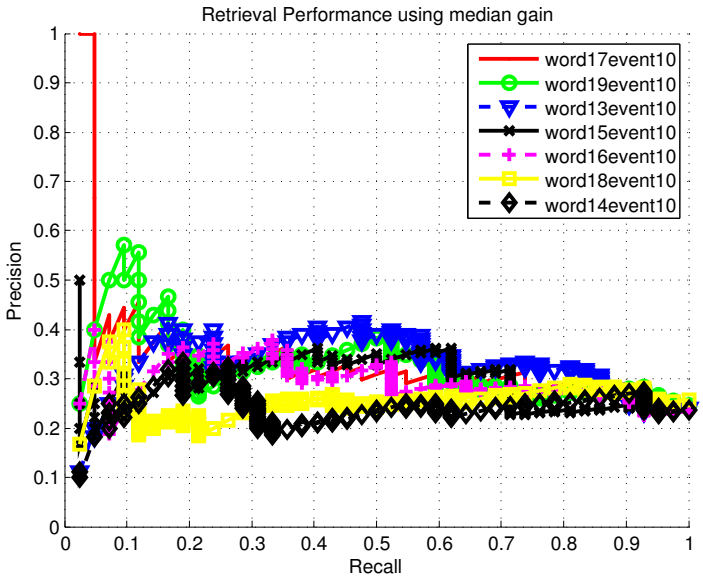
**Figure 26:** Precision-recall curve based on $max(G)$ for Video 1 in ChildPlay dataset.

1. The ranking criteria $max(G)$ and $mean(G)$ are better than $median(G)$ to evaluate the relevance of the retrieved items.

2. The average variations of APs are 15%, 10% and 6% for video 1, 2 and 3 respectively. Considering the fact that the parameters were varied up to 20%




(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$

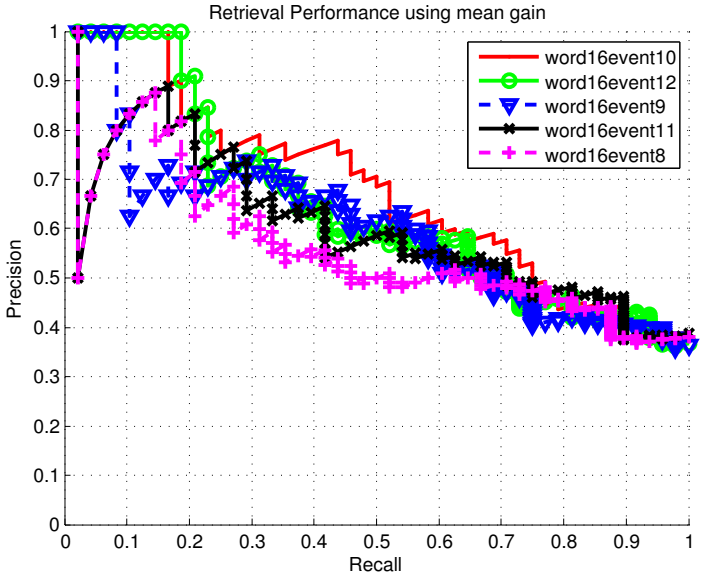



(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

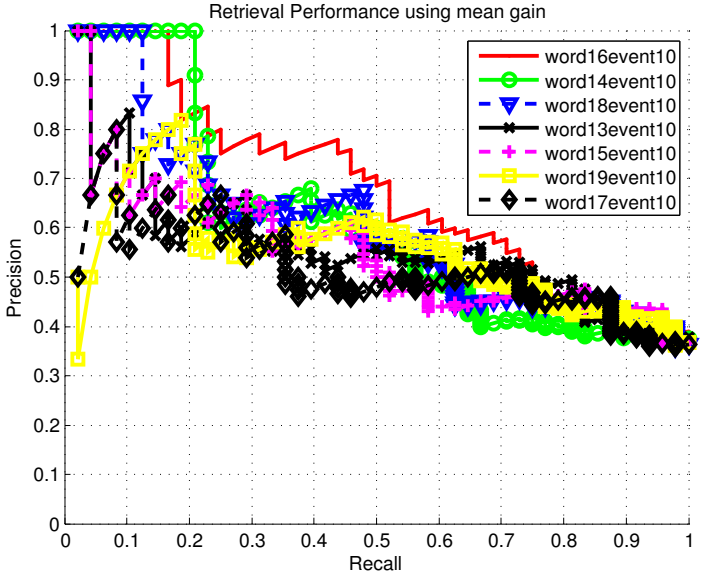
**Figure 27:** Precision-recall curve based on $median(G)$ for Video 1 in ChildPlay dataset.

with respect to the baseline choice $k_{event} = 10, k_{word} = 16$, the empirical study shows that our algorithm is robust to the change of the parameters.

3. The retrieval performance with different choices of parameters all demonstrates the promise of our method to effectively identify examples of social games in



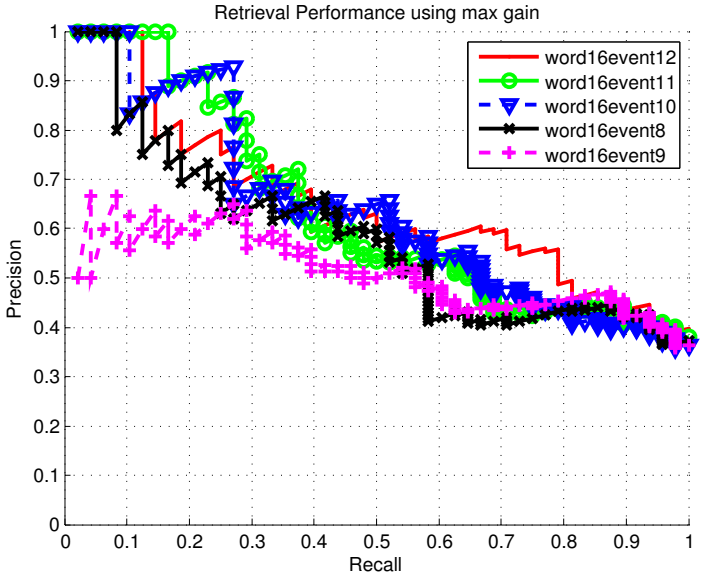

(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$



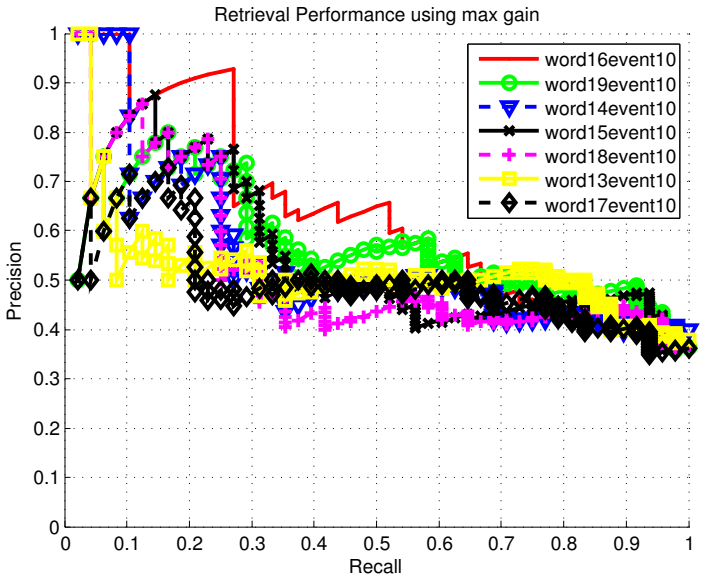

(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

**Figure 28:** Precision-recall curve based on $mean(G)$ for Video 2 in ChildPlay dataset.

unstructured videos. This is potentially useful for reducing the time and labor in the current practice of retrospective video study.




(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$

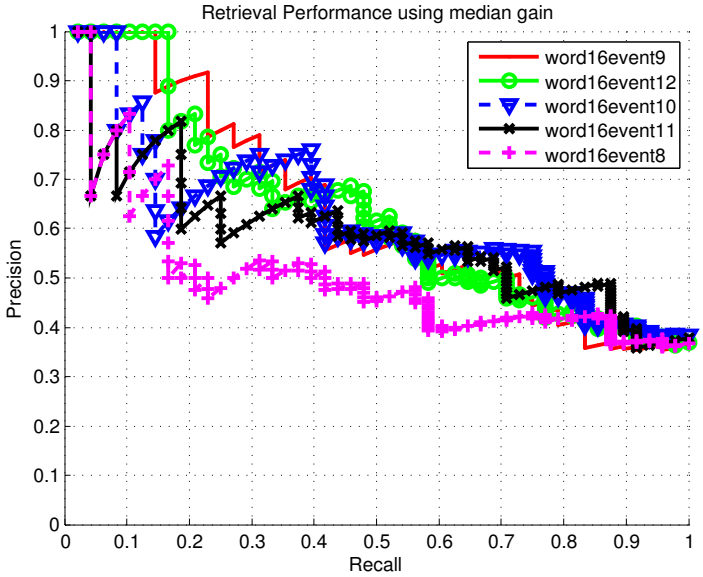



(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

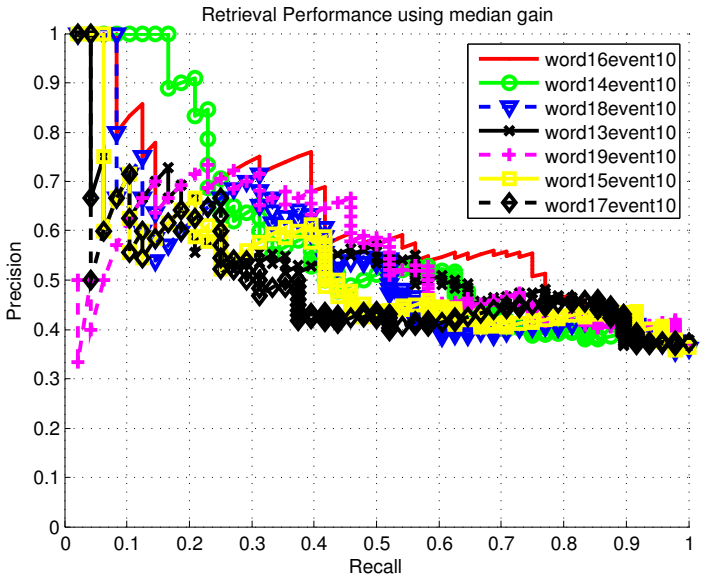
**Figure 29:** Precision-recall curve based on $max(G)$ for Video 2 in ChildPlay dataset.

## *4.5 ChildPlay: A Video Database of Children's Play*

We have collected a video database of realistic child play in natural settings, which we call "ChildPlay", along with the ground truth labeling of the social games. The dataset is published at our project website. The purpose of this video collection




(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$

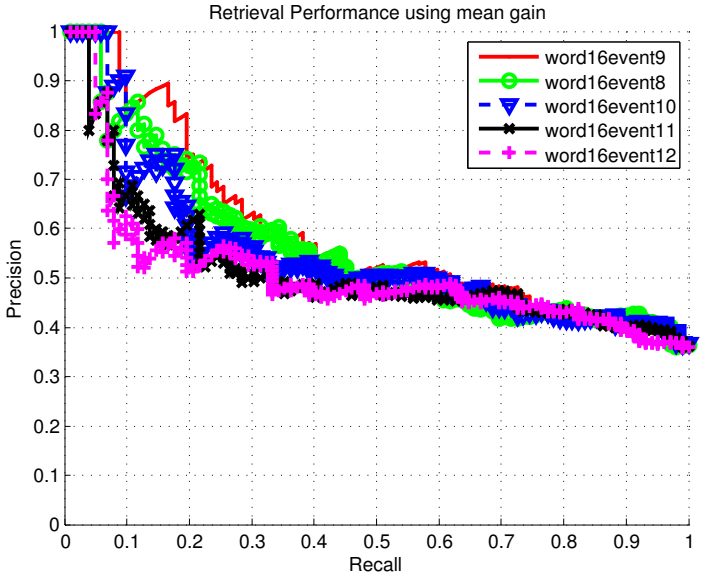



(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

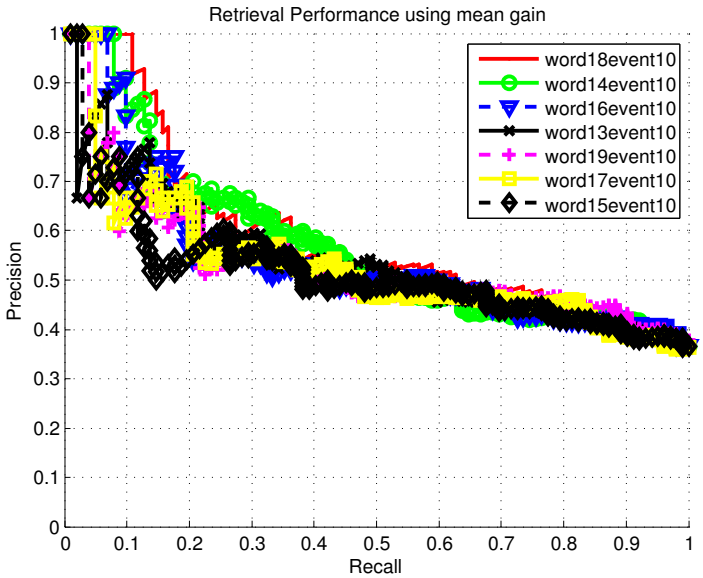
**Figure 30:** Precision-recall curve based on $median(G)$ for Video 2 in ChildPlay dataset.

is to support and encourage more research on human behavior analysis in realistic contexts, with the ultimate goal of understanding the abstract psychological/social states of the individuals from their behaviors.

The video database used for activity analysis can be classified into three categories:



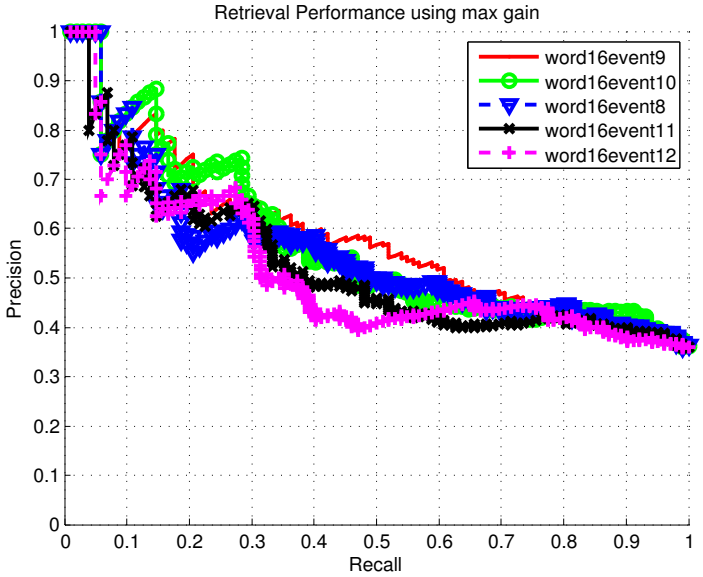

(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$



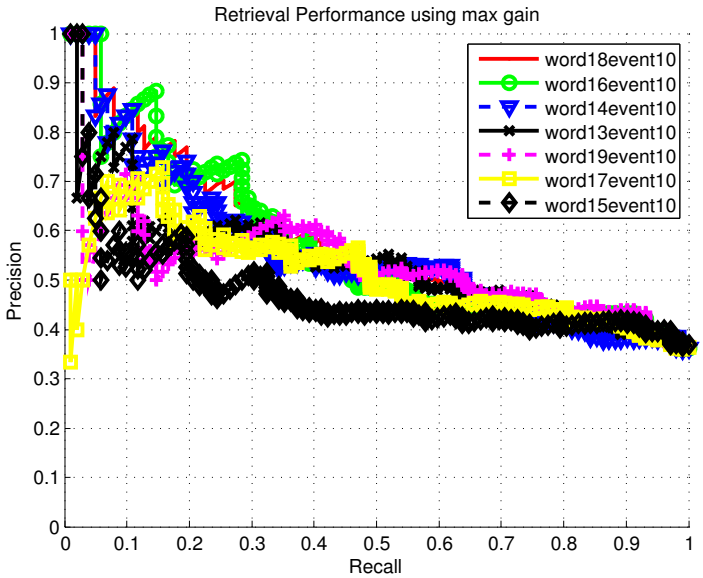

(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

**Figure 31:** Precision-recall curve based on $mean(G)$ for Video 3 in ChildPlay dataset.

surveillance videos, videos recorded in a structured setting and videos recorded in a natural setting. Activity analysis in surveillance videos is important for security applications. Examples include airborne videos on specific areas, videos recorded with static cameras in public places, such as an airport, train station, public parking



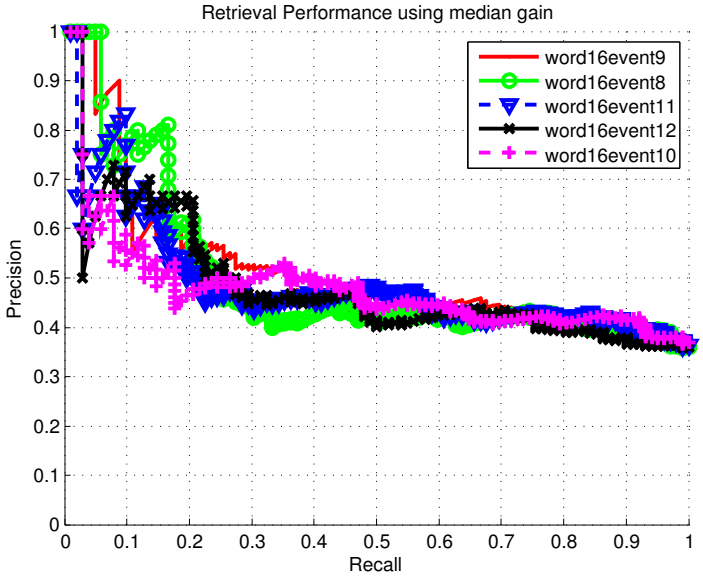

(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$



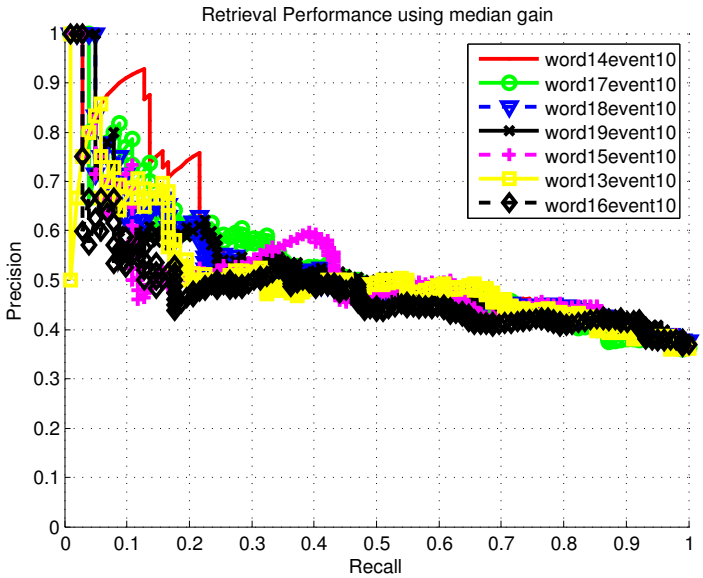

(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

**Figure 32:** Precision-recall curve based on $max(G)$ for Video 3 in ChildPlay dataset.

lots, or hospital. The videos recorded in a structured setting typically have at least one of the following properties: the performed actions are pre-specified, camera angle, motion and background are controlled by the researchers to support the capabilities and limitations of their activity analysis algorithms. Examples include the KTH




(a) $k_{word} = 16$,$k_{event} = 8, 9, 10, 11, 12$




(b) $k_{event} = 10$,$k_{word} = 13, 14, 15, 16, 17, 18, 19$

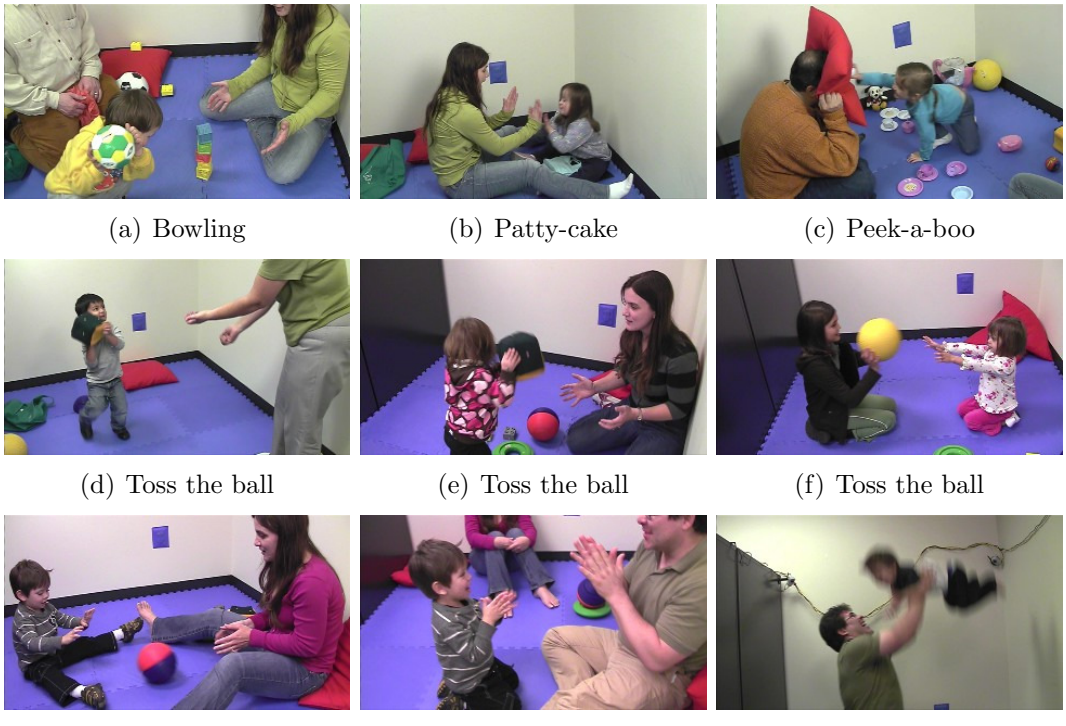
**Figure 33:** Precision-recall curve based on $median(G)$ for Video 3 in ChildPlay dataset.

dataset [75], articulated human motion data distributed by EHuM [22], and the videos collected by individual researchers for their own evaluation [9, 39, 44, 54, 78, 92]. Videos of human activity in a natural setting present significant challenges from a computer vision perspective. There exist two main sources for such videos: films and

home movies. A notable collection of film footage is from Laptev *et al.*'s work [42]. To our knowledge, our work is the first to use home movies for activity analysis.

Publicly available datasets have played an important role in computer vision history. Performance evaluation on the standard datasets has helped establish the state of the art in several areas. Examples include the Middlebury stereo vision comparison website [74], the face detection database [72], and the FERET evaluation methodology for face recognition algorithms [64]. Representative systematic evaluations on video analysis include PETS (IEEE workshop on Performance Evaluation of Tracking and Surveillance), EHuM [22], and TRECVID [84]. Our ChildPlay dataset is an important addition to the existing video collections and shall contribute to behavior understanding in natural environments.

### 4.5.1 ChildPlay video database description



(a) Bowling (b) Patty-cake (c) Peek-a-boo

(d) Toss the ball (e) Toss the ball (f) Toss the ball

(g) Roll the ball (h) Patty-cake (i) Toss baby

**Figure 34:** Examples of parent-child play.

The ChildPlay dataset features realistic activities in life, which contains a lot of

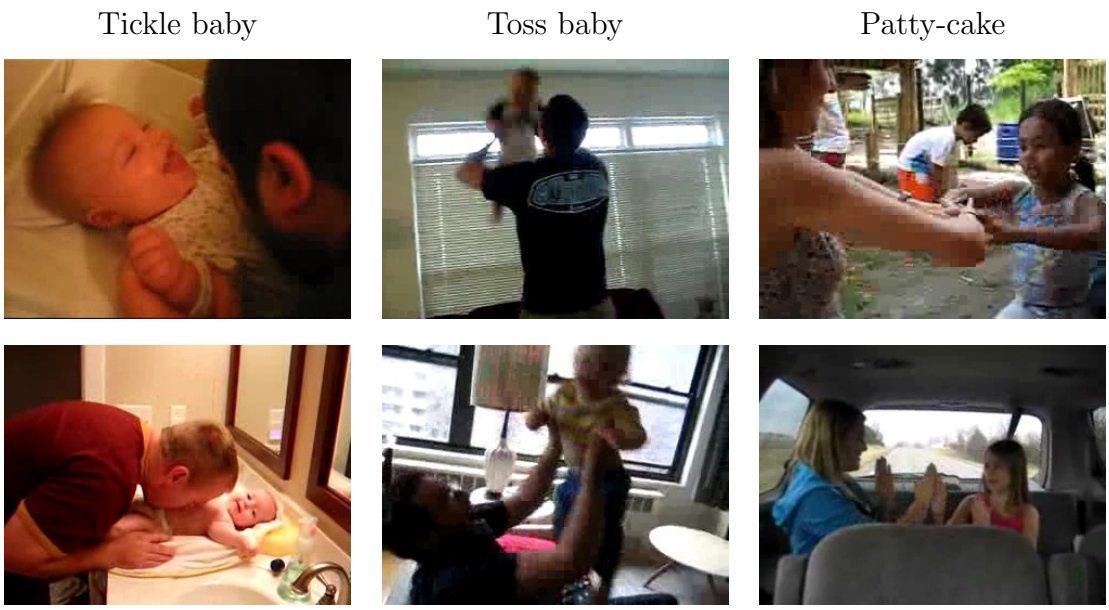



**Figure 35:** Examples of social games from YouTube.

social interactions. It includes parent-child play recorded in a laboratory setting, social games from YouTube, and home movies.

1. Parent-child play. We have collected 8 sessions of parent-child free play of 150 minutes total length. Figure 34 shows selected moments in this dataset. The videos were recorded in our child study lab at the Health Systems Institute (http://www.hsi.gatech.edu).

2. Social games from YouTube. A video only contains one game in this dataset. We have collected 10 peek-a-boos, 21 toss-the-baby, 44 patty-cake, and 25 tickle-the-baby. Figure 35 shows some examples.

3. Home movies. We have 1605 minutes of home movie that contains a lot of adult-child, child-child interactions.

## *4.6 Conclusions and Discussions*

Our vision of *Behavior Imaging* for helping autism research and clinical practice is to enable large-scale utilization of behavioral data by making it easier to collect such

data through the automated analysis of sensor streams (primarily vision and audio). Its goal is to develop integrated technologies for multi-modal computational sensing and modeling to capture, measure, and understand human behaviors. To support effective measurement and understanding of human behaviors, we have three sub-goals: retrieval technologies that allows efficient utilization of audio-visual recording of interactive behaviors, characterization of the interactions, and quantification of the behaviors. My work on social game retrieval is the initial attempt towards the goal of retrieval technologies.

Social games, as well as many other human interactions, exhibit a repetitive temporal structure with long temporal duration. Such a structure can be exploited for video analysis as an effective alternative to the conventional decomposition into actions. We have presented an unsupervised method for detecting quasi-periodic motion patterns that exploits the temporal structure of social games, and demonstrate its efficacy in parsing the stages of the social games, and retrieving social games and social interactions from unstructured video collections. The mined quasi-periodic patterns are sufficiently discriminative for categorization purpose. We present a support vector machine classifier based on the features extracted from the patterns in Chapter 5. The quasi-periodic pattern analysis also represents a substantial generalization of conventional periodic motion analysis. Chapter 6 demonstrates the generalization of our method at detecting motions of a range of quasi-periodicity. Compared with a periodic motion detector based on self-similarity [18], our approach demonstrates its superiority with better retrieval performance and the ability to extract periodic patterns of any length without the need for period estimation.

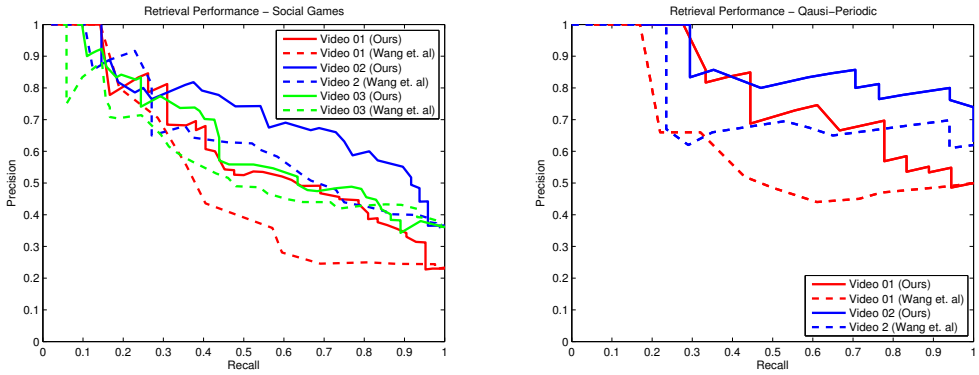
Our quasi-periodic pattern mining approach has its own limitations at retrieving social games. Social games consist of repetitions of turn-taking interactions between the dyad. Our current approach hasn't addressed the turn-taking interaction within each occurrence. By analyzing the current retrieval performance, we have shown that

**Figure 36:** A sequence of ball game in the ChildPlay dataset. Green indicates the causal set of visual words corresponding to the ball game interactions, and red indicates the causal set irrelevant to the ball game. These results were made by Karthir Prabhakar.

the false positives are mainly caused by the incomplete modeling of social games. Many retrieved videos contain repeated actions but not interactions. By analyzing the turn-taking interactions, we should be able to eliminate many false positives that do not exhibit interactive patterns among the visual words. For the false negatives, the main cause is that different frame labels are assigned to actions of similar purpose but have different poses. An example is given in Figure 12.

The sequential turn-taking interactions generate the temporal causal relationship between the visual words. A recent work by Prabhakar *et al.* shows improved retrieval performance on the ChildPlay dataset by analyzing the causalities between the visual words [67]. This work interpreted the sequences of visual words as a multivariate point-process. By using a spectral version of the pairwise test for Granger causality [30], we can group the visual words into independent causal sets. Figure 36 shows the causal set corresponding to the ball game (in green), and the other causal sets corresponding to background objects or camera motion (in red). Since camera motion, extraneous motions in the background can all lead to spurious visual words, by using the visual words from a meaningful causal group to construct the frame descriptors, and thereby to assign frame labels, we received better retrieval performance for both social games and the quasi-periodic events, which is shown in Figure 37.

To model the turn-taking interactions, one can extend the causal analysis to incorporate the spatial and temporal causalities among the visual words. I list five aspects that can be worked on to improve the current QP pattern mining approach. First, use scene detection instead of the sliding window approach to segment the




(a) Social games

(b) QP events

**Figure 37:** Retrieval performance on ChildPlay dataset. Solid lines depict the performance using causal set, and dashed lines depict the performance of QP pattern mining. These results were made by Karthir Prabhakar.

input video, since scene switching sometimes causes spurious interest points. Other camera stabilization techniques may also be used as a pre-processing. Second, automatically select the best $k_{event}$ and $k_{word}$. Third, allow the mining of approximately recurring patterns in a discrete sequence, so that we can tolerate that similar actions get assigned to different frame labels. Ukkonen [85] proposed a method for suffix tree based approximate string matching. Fourth, find recurring causal patterns from the visual words directly, instead of finding recurring patterns from the discrete sequence. The sets of words in different frames will have different sizes, and the challenge is to decide what the relevant words in each set are, and to extract the sequence of such sets that are (approximately) repeated. This approach should be considered together with the turn-taking interaction analysis. Minnen *et al.* proposed a method for detecting subdimensional motifs shared by multiple time series data [53]. Their method is not suitable for the retrieval task, as the repetitions in games have different lengths and their method looks at recurring motifs from sequences of the same length. The method first searches for motif seeds from pairs of similar, fixed-length subsequences, and then uses these seeds to detect other occurrence of the same motif from other sequences. Fifth, incorporate supervised detections to our pattern mining framework.

For example, we can detect the individual actions, and find a recurring sequence of the detected actions. Such recurring patterns can be used to predict the next action. On the contrary, we may predict the type of game and the object (if any) that is used in the game simultaneously based on the interactive patterns of the visual words. For instance, a ball can be tossed or rolled in a ball game. Different ways of playing will have different motion patterns in the videos, which can be leveraged to predict the game type and the object in use.

The current retrieval performance has a significant space for improvement. One of our ultimate goal is to build a tool for psychologists to support automatic video content filtering. In behavioral study, it is important for psychologists to retrieve all the behaviors of interest, under any situation (*e.g.*, at home, at school, or outdoor), so that they can explore the trajectories of behaviors in different situations. We need to reduce false positives when recall reaches 100%. From the above proposed improvements for retrieval, the most important direction is to incorporating the turn-taking interaction into the pattern mining framework. By analyzing the turn-taking interactions, we can not only eliminate some of the false positives, but also get prepared to automate the quantitative measurement of the interactions.

We conclude that video collections of parent-child interaction constitute a rich source of behavioral data which is useful to psychologists and amenable to computer vision analysis. By publishing our video dataset and annotations at the project website, we hope more progress can be made in this exciting new domain of behavior understanding.

# CHAPTER V

# CATEGORIZATION OF SOCIAL GAMES BASED ON QUASI-PERIODIC EVENT ANALYSIS

The quasi-periodic pattern mining approach presented in Chapter 4 successfully extracts the patterns corresponding to meaningful stages of the social games from real-world videos. It suggests that meaningful analysis of extended interactions can be obtained by considering the temporal structure of the video represented by the visual words, without a decomposition into specific actions. Since social games are governed by the underlying abstract game rules, it is worth asking whether the QP patterns represent the game rule and how discriminant the patterns are at categorizing games?

We examine the discriminancy of the QP patterns by building a social game classifier based on the QP pattern analysis. Automatic social game categorization is beneficial to behavioral study. For example, statistical information about the type of social interaction gestures and their frequencies in videos is often collected for behavior assessment [16, 60, 71, 81].

The problem of classifying social games in unstructured videos is very challenging, due to their two properties of *quasi-periodicity* and *multi-instantiation*. Quasi-periodicity describes two facts. First, the length of every repetition of the game and the duration of each stage may vary since the parents often change the rhythm of the game to keep the babies engaged and to avoid overstimulating them. Figure 7 shows two occurrences of the hide-and-reappear sequence in a peek-a-boo game. One takes approximately 15 frames to completely hide the toy (Figure 7(a) to 7(c)), and the other takes 44 frames (Figure 7(f) to 7(h)). The second aspect of quasi-periodicity is that random actions can be inserted or deleted during play, such as kissing the

baby while playing peek-a-boo with him. Typical bag-of-words (BOW) model based approaches often extract feature representations from a fixed-length space-time volume, or at multiple scales [42, 45, 58]. This is a simple solution for actions spanning a short time-scale window. For extended complex interactions, such as the quasi-periodic social games, a compact and effective BOW representation needs to include the characteristic space-time interest points (STIPs) within the volume, and not to include the STIPs generated by random actions. The multi-instantiation property describes the fact that a game can be played in many different ways as long as the underlying abstract game rule is followed. Consequently, it is difficult to collect a representative video database to support supervised learning of game strategies. The QP patterns provide an alternative view of game strategies, and it is possible to construct a representative space based on the patterns.

We address the challenging social game categorization problem by leveraging the work on quasi-periodic game analysis. Our method automatically selects the most relevant STIPs in constructing the feature representation and the visual codebook for learning purpose. We utilize the QP pattern analysis in two aspects. First, we use the representation of QP patterns as the feature representation for game categorization. This is based on the hypothesis that if the mined patterns do in fact correspond to meaningful game stages, then feature representations built from these patterns should provide an effective basis for categorization. In particular, if mined patterns are representative of the abstract game rule, they should be shared by all the instances of the same game. Second, we retrieve the STIPs that belong to the correspondent frames of the QP patterns, and use these characteristic STIPs to build the codebook for learning purpose. Our experiments validate this assumption and demonstrate very promising performance on classifying social games collected from YouTube. In addition, we show that our method can also be used to categorize videos of sports rallies (collected from YouTube by Karthir Prabhakar), demonstrating the generality

of our approach.

## *5.1 Approach*

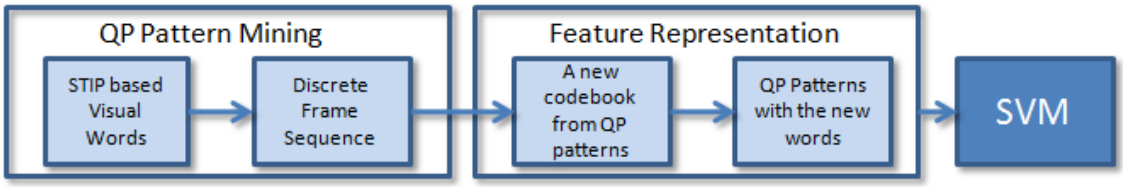
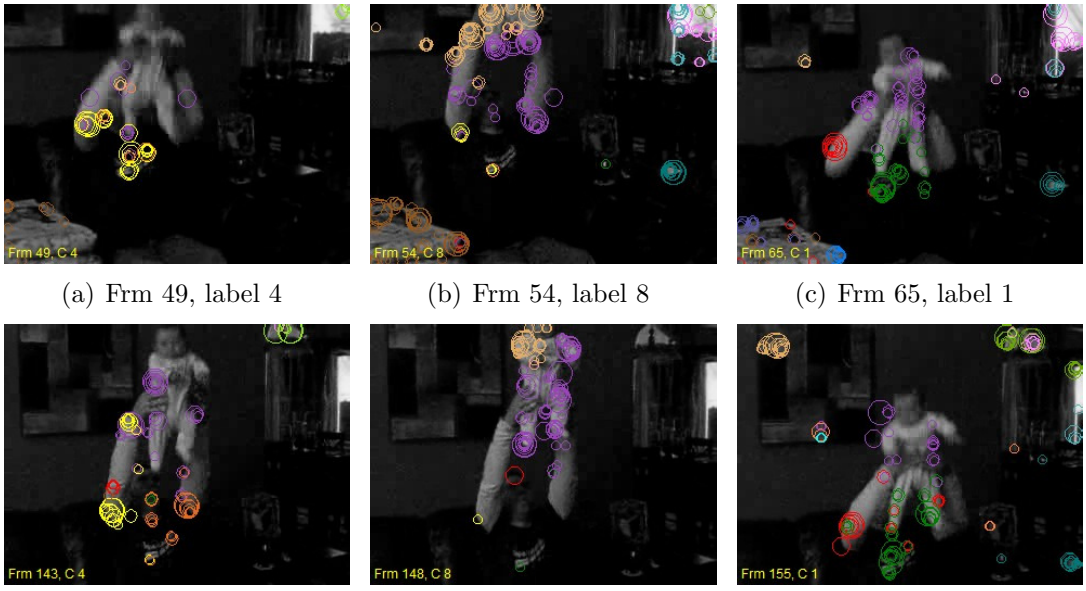



**Figure 38:** Pipeline of our categorization approach.

This section presents our approach for classifying recurring interactions in social games. Figure 38 shows the pipeline of our approach. This work is built upon the QP pattern analysis presented in Chapter 4. We demonstrate the generality of our approach by evaluating it on multiple datasets: social games and sports rallies collected from YouTube. All the videos were first segmented into clips of length $t\_win$ using a sliding window. $t\_win$ was long enough to cover at least two occurrences of the quasi-periodic activities. Our goal is to learn the structures of the interactions from the training clips, and predict category labels for the testing clips.

### 5.1.1 Quasi-periodic pattern extraction

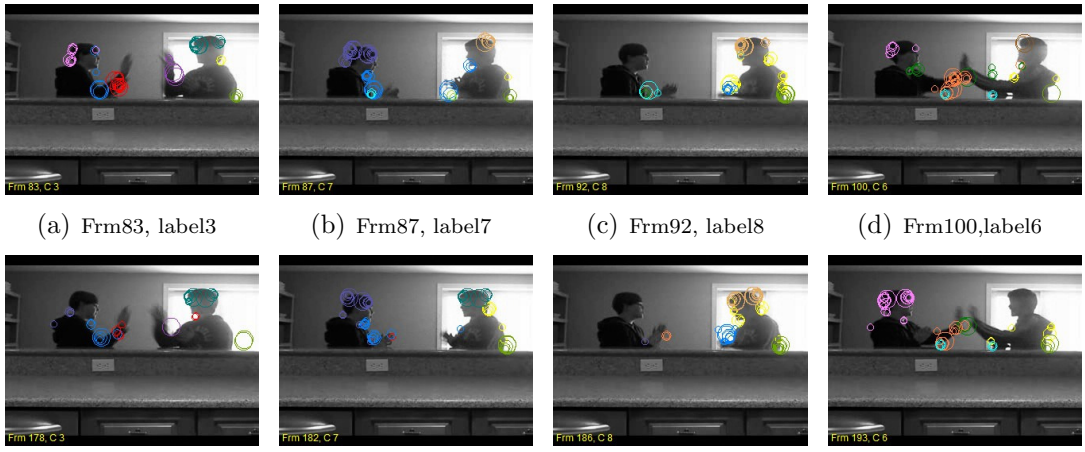
Every clip was processed independently to find the QP patterns using our QP pattern mining approach [86] (see Chapter 4 for details). A video clip of recurring activities often contain many QP patterns and a pattern can occur several times in the sequence. The mined patterns frequently correspond to meaningful stages of a game. Figure 39 shows two occurrences of a mined pattern 4-8-1 from a toss-baby clip. Label 4 corresponds to the action of the father starts to lift the baby. Label 8 is when the baby is lifted to the highest position. Label 1 is when the father holds the baby back down. Pattern 3-7-8-6 shown in Figure 40 describes a sequence from retrieving hands to clapping their left hands in a patty-cake game. Figure 41 shows an example pattern mined from a tennis match video.




(a) Frm 49, label 4 (b) Frm 54, label 8 (c) Frm 65, label 1

(d) Frm 143, label 4 (e) Frm 148, label 8 (f) Frm 155, label 1

**Figure 39:** Mined pattern 4-8-1 and its two occurrences from a toss-baby clip. The interest points are color-coded according to visual word assignments.




(a) Frm83, label3 (b) Frm87, label7 (c) Frm92, label8 (d) Frm100,label6

(e) Frm178,label3 (f) Frm182,label7 (g) Frm186,label8 (h) Frm193,label6

**Figure 40:** Mined pattern 3-7-8-6 and its two occurrences from a patty-cake clip.

### 5.1.2 Feature representation from quasi-periodic patterns

After extracting the QP patterns, we collected the STIPs belonging to the correspondent frames of the mined patterns from all the training videos across categories to build a new visual vocabulary. This interest point set $P'$ was a subset of the original interest point set $P$. Each $p \in P'$ was represented by its HOF descriptor, as described

(a) Frm835, label6 (b) Frm839, label5 (c) Frm843, label4

(d) Frm922, label6 (e) Frm926, label5 (f) Frm933, label4

**Figure 41:** Mined pattern 6-5-4 and its two occurrences from a tennis clip.

in Sec. 4.1 without the normalized spatial coordinates $(x, y)$, resulting in a feature vector of 144 dimensions. The visual vocabulary was then generated by k-means clustering of all the feature vectors with $k = 30$.

When extracting the QP patterns, we let the pattern score threshold adapt to the entire pattern scores in that video. Denote the set of all the pattern from a clip as $QuasiPatSet$, and pattern score $G = G(Pat), Pat \in QuasiPatSet$. The threshold was set as $max(mean(G), median(G))$ for that clip.

With the new codebook, each selected STIP was assigned to its nearest visual word based on the Euclidean distance between their feature vectors. Every occurrence of the mined pattern produces a data example, which is the histogram of the visual words in that pattern occurrence. So each feature datum is a 30-dimensional vector.

Our method of generating features has three advantages. First, a training clip typically generates many training examples, because multiple patterns are mined and each of them occurs at least twice. This is an effective way to boost the training set. Second, our features are build from STIPs that characterize the recurring activities.

This is important when classifying unstructured videos such as home movies and YouTube videos, in which many spurious STIPs could be detected due to camera motion, shaking and background motions. The pattern mining process hopefully eliminates a lot of these points as they are less likely to occur repetitively and have high pattern scores. Nevertheless, our training data is still noisy as we are looking at complex human interactions in real-world footage. Third, the labelled training data is obtained without any human labor.

### 5.1.3 Non-linear Support Vector Machine Classifier

We train a non-linear support vector machine (SVM) [15] using a Gaussian kernel with $\chi^2$ distance [42]. The kernel is defined as:

$$K(h_i, h_j) = exp(-D_\chi(h_i, h_j)/c) \tag{8}$$

$$D_\chi = 0.5 * \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}} \tag{9}$$

where $h_i = \{h_{in}\}$ and $h_j = \{h_{jn}\}$ are feature data $i$ and $j$. $n$ is the bin index. V is the vocabulary size. $V = 30$ in our experiments. $c = mean(D_\chi)$.

### 5.1.4 Voting scheme for clip classification

The SVM classifier is trained to make predictions on every occurrence of the mined patterns from the testing clip. In order to predict class label for every testing clip, individual predictions on the occurrences need to be consolidated. We used a winner-take-all scheme to combine the individual votes — the class label for the video clip was determined as the dominant label among the mined patterns, while a pattern's label was the dominant label among all its occurrences.

## *5.2 Experimental Results*

We evaluate our method on two challenging video datasets from YouTube: social games and sports rallies. Both datasets contain repetitive interactions that can be

extracted with our QP pattern mining method [86]. For the social game dataset, $t\_win = 500$ on average. For the sports dataset, $t\_win = 2000$. We empirically chose a long time-scale window for the sports videos for two reasons: first, the players in a match play stop frequently to prepare for the next turn, which results in time intervals of few motions; second, the videos views often switch to the audience for a while during a match, which does not necessarily result in QP motion patterns.

We also compare the categorization performance of our method with a baseline approach — constructing codebook from a random subset of all the interest points [42]. To make fair comparison, the same number of STIPs were used to construct the codebook. The classification performance shows that our method consistently achieves higher categorization accuracy than the baseline approach.

**Social Game Dataset**: This dataset contains three games: toss-baby, patty-cake and tickle-baby. Table 10 summarizes the number of videos, the number of clips per category, and the size of the training and testing data. The classification performance is measured with confusion matrix, shown in Figure 42. Our method significantly outperforms the baseline approach on this highly diverse dataset. Figure 43 is a stack view of the predictions of the individual patterns from each video clip. Figure 44 shows some examples of correct and incorrect classifications.

**Table 10:** Statistics on YouTube Social Game Dataset.

| | Toss-baby | Patty-cake | Tickle-baby |
|---|---|---|---|
| #videos | 21 | 43 | 25 |
| #clips | 38 | 180 | 76 |
| #Training clips | 26 | 59 | 51 |
| #Training examples | 766 | 1856 | 1709 |
| #Testing clips | 12 | 121 | 25 |
| #Testing examples | 331 | 3668 | 630 |

**Sports Video Dataset**: This dataset contains three sport rallies: tennis, beach volleyball (volley) and table tennis (TT). Table 11 summaries all the statistics about the dataset. The confusion matrices for our method and the baseline approach are

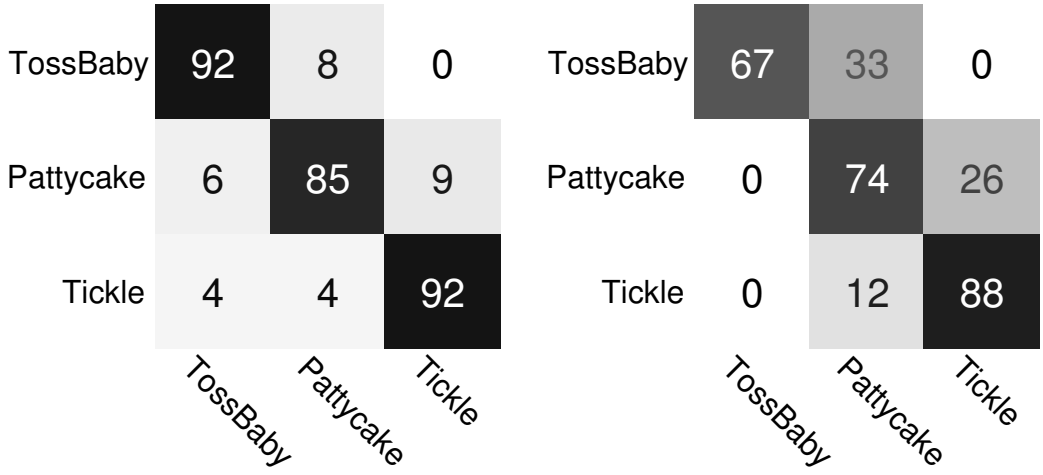
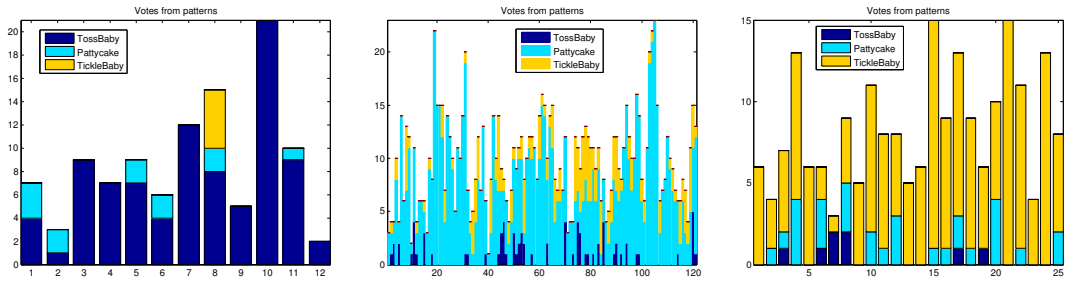



**Figure 42:** Confusion matrices (in percentage) on YouTube Games. Left: our approach. Right: baseline.




(a) Toss-baby (b) Patty-cake (c) Tickle-baby

**Figure 43:** Individual pattern's vote per category. x-axis: index of clips. y-axis: count of votes. Blue: toss-baby. Cyan: patty-cake. Yellow: tickle-baby.

shown in Figure 45. For this dataset, our approach performances slightly better than the baseline on TT, and is on pars with the baseline on the other two sports. This is probably because the dataset consists of clips from TV programs, which share some common editing formats, and these views themselves could be characteristic of the sports games. For example, the camera often switched to a similar closeup view of the player after each scoring, then switched back to the full-view of both players. Figure 46 shows four most common views in a table tennis video: a distance view, close-up views of the individual player, and a view of the audience. Figure 47 shows the distribution of individual pattern's vote for the clip category prediction. Selected correct and incorrect predictions are shown in Figure 48.


**Figure 44:** Selected correct (top) and incorrect (bottom) predictions for the YouTube social game dataset.

**Table 11:** Statistics on Sports Rally Dataset.

| | Tennis | Volley | TT |
|---|---|---|---|
| #videos | 2 | 5 | 5 |
| #clips | 28 | 34 | 48 |
| #Training clips | 20 | 13 | 20 |
| #Training examples | 991 | 1176 | 839 |
| #Testing clips | 8 | 21 | 28 |
| #Testing examples | 361 | 1320 | 1645 |

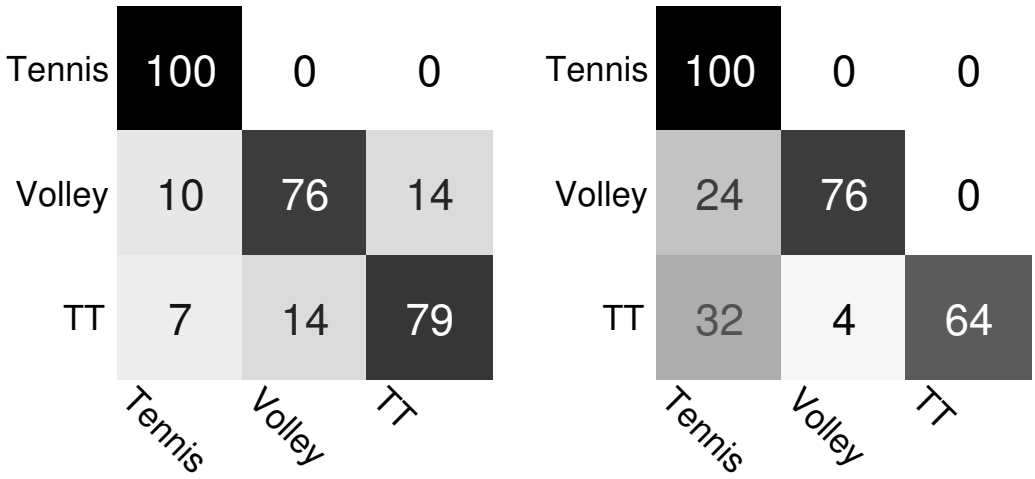
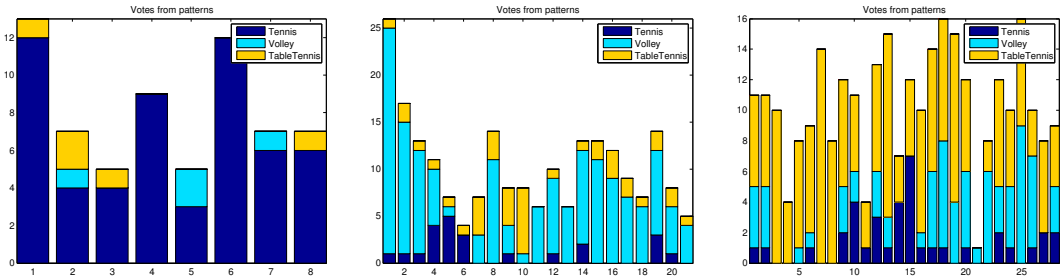



**Figure 45:** Confusion matrices (in percentage) on sports videos. Left: our approach. Right: baseline.

## *5.3 Conclusions*

Repetitiveness provides a strong cue for behavior analysis in complex videos. We have presented an approach of selecting informative space-time interest points for the categorization of extended human interactions with repetitive temporal structure


**Figure 46:** Four common views in a table tennis match.




(a) Tennis (b) Volley (c) Table Tennis

**Figure 47:** Individual pattern's vote per category. Blue: tennis. Cyan: volley. Yellow: table tennis.


**Figure 48:** Examples of correctly predicted (top) and incorrectly predicted (bottom) videos in sports dataset. All Tennis videos are correctly classified.

in complex real-world footage. We leverages our previous work on quasi-periodic pattern mining for social game retrieval [86] to extract the STIPs that correspond to the meaningful game stages, and to build feature representation for classification with a non-linear support vector machine classifier. We have also assembled two novel interaction action datasets from YouTube: parent-child social games and sports rallies. Our method consistently outperforms a baseline approach which selects STIPs by random sampling.

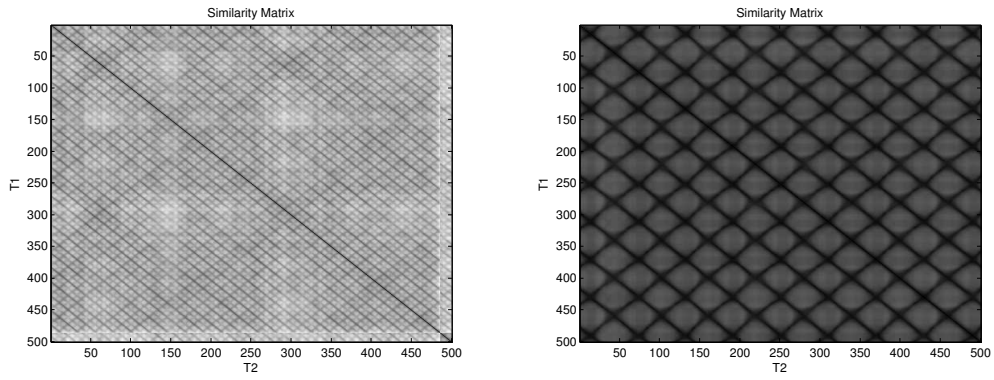
Enforcing the temporal ordering [59] or the structural information [94] of visual words is shown to increase the discriminative power of the sparse STIP representation. We plan to incorporate the spatio-temporal structure of the selected STIPs to improve the classification performance. Motivated by the real problems in behavioral science, the development of automated methods for social interaction retrieval and categorization will usher in a new area of quantitative behavior understanding in computer vision.

# CHAPTER VI

# AN EMPIRICAL CHARACTERIZATION OF QUASI-PERIODIC MOTIONS

Our quasi-periodic event analysis described in Chapter 4 represents a substantial generalization of conventional periodic motion analysis. For example, periodic motion is a special case of quasi-periodic events. In this chapter, we will first analyze a representative periodic motion analysis method based on self-similarity measurement [18]; then we give our definition of quasi-periodicity, and illustrate the elements of quasi-periodicity with video examples and the corresponding self-similarity plots; finally we will compare and analyze the performance of retrieving motions over a range of quasi-periodicity using our method and Cutler and Davis's approach [18].

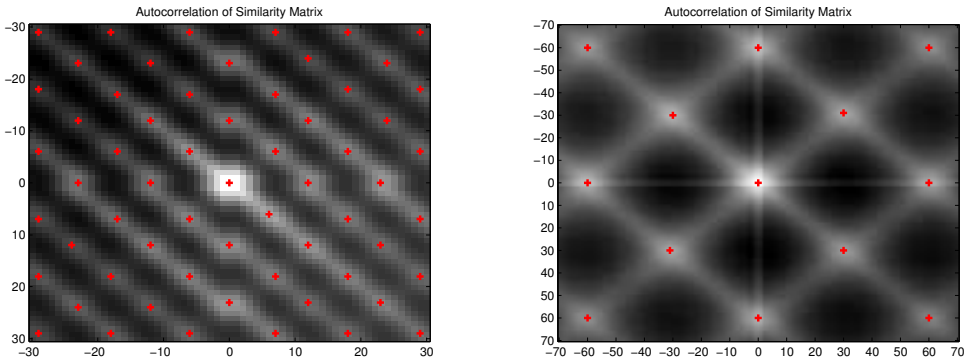
## *6.1 Periodic Motion Analysis based on Self-Similarities*




(a) Running (b) Rowing

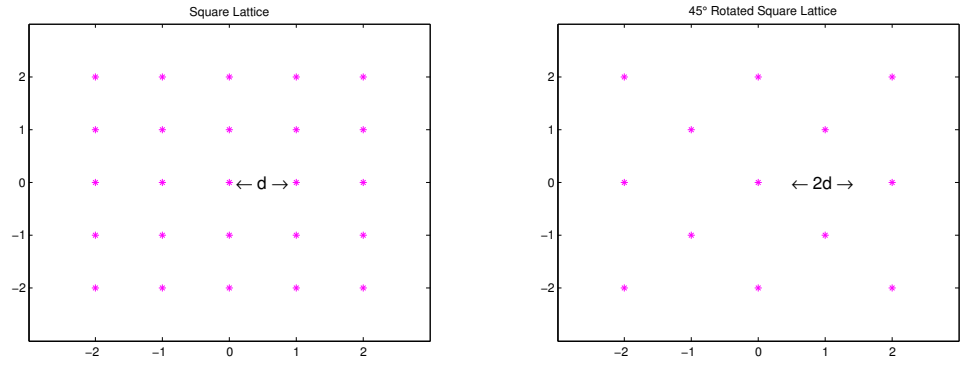
**Figure 49:** Similarity matrices.

Periodic motion can be mathematically characterized by Equation 10, where $x$




(a) Running (b) Rowing

**Figure 50:** Normalized autocorrelation of the similarity matrices in Figure 49. Peaks are denoted with red +.




(a) Square lattice (b) 45° rotated square lattice

**Figure 51:** Two lattices used for periodic motion analysis based on self-similarities [18].

represents the object's periodic motion signal, $T$ is the period, and $n = 0, 1, 2, \ldots$.

$$x(t) = x(t + nT) \tag{10}$$

Suppose the periodic motion does not have any noise and perturbation, the distance between the motion at any two moments that are separated by $nT$ should be zero. In other words, $\text{dist}(x(t), x(t + nT)) = 0$. For most periodic human motions, such as running and walking, the poses can vary from time to time, but the overall variations should be small at the correspondent periodic intervals. This is the foundation of Cutler and Davis's approach for periodic motion analysis [18]. They made

the observation that periodic motions in videos lead to periodic similarities between pairs of frames of a certain distance (usually the period $T$ or $nT$), and robust periodic motion analysis can be conducted on the similarity matrix $S$.

Any metric can be used to measure the similarity between the moving object at two frames. We followed Cutler and Davis's choice of absolute correlation between any two frames.

$$S(i,j) = \sum_{(x,y)\in O_i} |O_i(x,y) - O_j(x,y)| \tag{11}$$

where $O_i$ and $O_j$ are the moving object at frame $i$ and $j$ respectively.[1] In our experiment, we use the entire frame for self-similarity calculation instead of the moving person, since the person is not changing size, and the background is mostly stable and homogeneous. Figure 49 shows two similarity plots of a person running on a treadmill (Figure 55(a)) and rowing with a rowing machine (Figure 55(b)) respectively. The image intensity is scaled to [0,255]. The darker regions indicate higher similarity between frame $T1$ and $T2$. The diagonal is always dark as an image is always similar to itself at any time. These two periodic motions generate similarity plots that have dark lines parallel to the diagonals. The dark lines perpendicular to the diagonal are caused by the similarity between $x(t)$ and $x(T-t)$.

Periodicity can be detected by estimating the power spectrum on the columns of $S$, assuming that they are stationary and contaminated with white noise. The peaks in the spectrum indicate the existence of periodic motions in the sequence. However, the stationary assumption does not always hold in realistic human motion. Short-time Fourier Transform (STFT) can be used to capture local periodicity. The short-time analysis window is chosen to be equal to several periods in practice. STFT suffers when significant non-Gaussian noise exist in the columns of $S$, or if the period is not locally constant.

[1] Minimal $S$ can be searched within a local region for every $(x,y)$ to account for tracking/segmentaiton errors.

Cutler and David proposed an alternative robust method to detect the periodicity from the similarity matrix [18], which is to analyze the structure of local peaks of the autocorrelation matrix of $S$. The normalized autocorrelation of $S$ is defined as:

$$A(d_x, d_y) = \frac{\sum_{(x,y)}(S(x,y) - \bar{S})(S(x+d_x, y+d_y) - \bar{S_L})}{(\sum_{(x,y)}(S(x,y) - \bar{S})^2 \sum_{(x,y)}(S(x+d_x, y+d_y) - \bar{S_L})^2)^{0.5}} \quad (12)$$

where $S_L$ is $S$ shifted by the lag $(d_x, d_y)$ (zeros are padded when necessary), $\bar{S}$ is the mean of $S$, and $\bar{S_L}$ is the mean of $S_L$. Figure 50 shows the normalized autocorrelation matrices corresponding to the similarity plots in Figure 49.

Periodic motions always lead to distinguished lattice structures of the local peaks in $A$. The periodicity in motion is analyzed via the matching of the peaks $P$ to the lattice structure. Two types of lattice structures are studied (Figure 51): square and 45° rotated square. The rotated square is a subcategory of square lattice and it is generated by the motions that not only are periodic with period $T$, but also are similar at $t$ and $T - t$. A typical example is the side view of a person walking or running.

The matching process has two steps. Denote the peaks in the square lattice as $M_{S,d}$ and the peaks in the rotated square lattice as $M_{R,d}$. First, the closest set of peaks $B_i = \{P_i|, P_i \in P\}$ to $M_{d,i}$ are found, which satisfies the following three conditions:

$$\{P_i||M_{d,i} - P_i| \leq \min_{j \neq i} |M_{d,i} - P_j|\} \quad (13)$$

$$\{P_i||M_{d,i} - P_i| \leq T_D\} \quad (14)$$

$$\{P_i|A(P_i) \geq T_A\} \quad (15)$$

where $T_D$ is the maximum distance $P_i$ can deviate from $M_{d,i}$, $T_A$ is the minimum autocorrelation value that $B_i$ should have. In our experiments, we set $T_D = d/2$, $T_A = 0.25$.

Second, $P$ matches $M_d$ if the following two criteria are met:

$$\min_{d_1 \leq d \leq d_2} e(M_d) < T_{avge}, e(M_d) = \frac{\sum_i |M_{d,i} - B_i|}{|B|} \quad (16)$$

$$|B| \geq T_{ratio}|M_d| \tag{17}$$

where $e(M_d)$ is the average distance between matched lattice points the local peaks from $A$. [2] $T_{ratio}$ is the percentage of matched lattice points in the lattice structure. In the experiments, we set $T_{avge} = d/2$, $T_{ratio} = 0.3$. The best $d$ is searched over the range of $[d_1, d_2]$ to find the best matching lattice structure.

## *6.2 Elements of Quasi-Periodicity*

We describe any motion or activity as quasi-periodic if it consists of repetitions of motions or activities, with a range of permissable variations that can be recognized by human perception.

From this definition, we derive the following elements of quasi-periodicity, which can be illustrated by our video dataset:

- varying length of period $T$ at different repetitions;
- random insertion or deletion of actions; and
- varying poses for the same action.

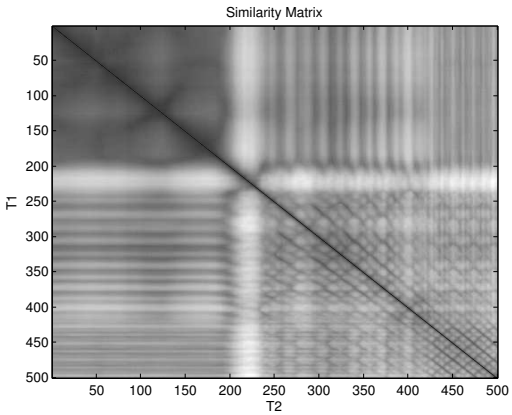
Figure 52 shows an example of quasi-periodic motion with elements of varying $T$ and random inserted action. In this video clip, a person first made a stop on the treadmill to drink some water (Figure 52(a)), then adjusted her running choices (Figure 52(b)), and resume running in the end (Figure 52(c)). In the similarity plot, we can see the part corresponding to the resting state, where similarity score is high for every frame since the person didn't have much motion; and the part corresponding to the running at the end where there are dark curves parallel to the diagonal. Notice the distance between the diagonal curves are getting smaller and smaller because the person was speeding up.

[2] $e(M_d) = \sum_i |M_{d,i} - B_i|$ was used in [18], but we found the average error has better retrieval performance in our experiments.

(a) Frm 1

(b) Frm 72

(c) Frm 268



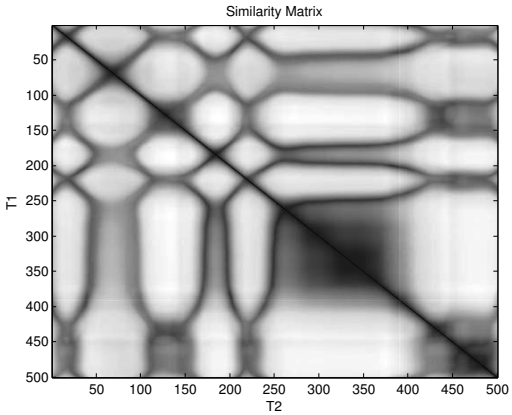

(d)

**Figure 52:** Selected frames from a running sequence. (a) stop and drink water. (b) select running options on the panel. (c) start running. (d) similarity plot.

(a) Frm 50

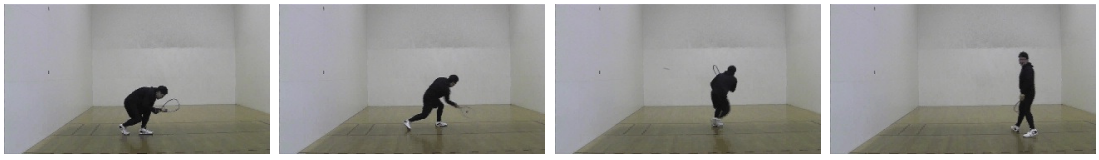
(b) Frm 132

(c) Frm 181

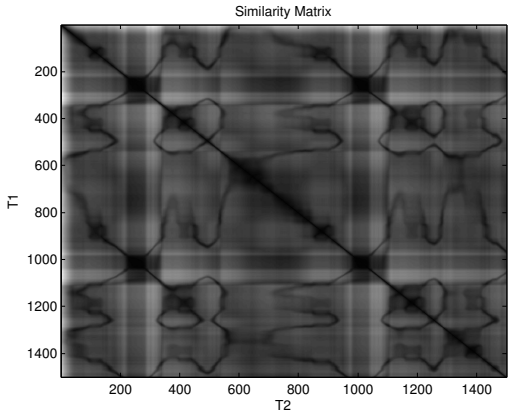



(d)

**Figure 53:** Selected frames from a rowing sequence. (a) rowing. (b) stop and take the bottle. (c) drink while rowing back and forth. (d) similarity plot.



(a) Frm 103 (b) Frm 144 (c) Frm 154 (d) Frm 180




(e) Similarity plot

**Figure 54:** Selected frames from a racquetball serve practice.

Another example shown in Figure 53 has all the three elements of quasi-periodicity. At the beginning of the sequence, the person was doing rowing exercise (Figure 53(a)). Then she stopped to take her bottle (Figure 53(b)), and moved back and forth on the rowing machine while drinking water for a while (Figure 53(c)). She pulled up the handle and was about to row again in the end of the sequence. The similarity plot shown in Figure 53(d) demonstrates the varying $T$ clearly. The varying poses are demonstrated by Figure 53(a) and 53(c). Although the upper body had different poses, the legs were doing similar back and forth movements, which leads to dark regions around position (50,181) in the similarity plot.

In principle, the periodic motions or activities should include the case where a single period contains a series of actions that occur over a time-scale of tens of seconds or longer. But most video-based periodic motion analysis only focus on short time-scale motions such as walking or running [18, 66]. We show an example of quasi-periodic motion that each period consists of a series of short actions in Figure 54. The

sequence is a practice of racquet ball serves played by a non-professional. A complete period of a serve practice includes prepare, hit, pick up the ball and get ready for the next serve. We can see dark curves parallel to the diagonal that correspond to recurrences of the practice in the similarity plot (Figure 54(e)), but the overall structure is highly irregular due to the variations of the speed and gestures during practice. Therefore it is difficult to detect such quasi-periodicity using conventional periodic motion analysis methods.

## *6.3 Experimental Evaluation*

In this section, we compare and analyze the performance of retrieving motions/activities over a range of quasi-periodicity using our method and the self-similarity approach [18]. For both methods, the retrieval is done by detecting (quasi-)periodic motions in each sliding window, same as the framework used in Chapter 4. We assume the window length $t\_win$ is long enough to contain at least two occurrences of the repetitive actions. We set $t\_win = 500$ on average in our experiments. For the racquetball serve practice, we set $t\_win = 1500$ since one cycle of serve takes about 350 frames.

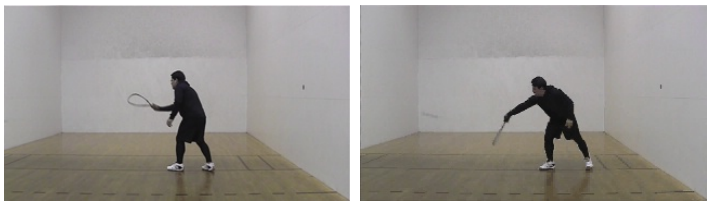
When retrieved with Cutler's method, the structure of the local peaks in the autocorrelation matrix $A$ is examined and matched to the two lattice structures. For a video segment, if for all the possible lattice structures that are compared with, the number of matched local peaks $|B| < T_{ratio}|M_d|$, we label the video as not containing periodic motions. The precision-recall curve is generated by calculating the precision and recall at each position in the retrieved items ranked according to its $e_(M_d)$. When retrieved with our method, three precision-recall curves are generated, corresponding to the three measurements of the mined $QuasiPatSet$ scores: $mean(G)$, $max(G)$ and $median(G)$, where $\{G = G(Pat)|Pat \in QuasiPatSet\}$.

Unlike self-similarity approach, which needs to choose a set of parameters and is designed to analyze periodic motion only, our method does not require any knowledge

of period length, and how and where the quasi-periodicity occurs. Consequently, our method should have better retrieval performance. In addition, the retrieved quasi-periodic patterns often parse the period into meaningful stages.

### 6.3.1 Video dataset description

(a) Running (b) Rowing (c) Stretching



(d) Forehand practice (e) Backhand practice

**Figure 55:** Selected periodic and quasi-periodic motions.

**Table 12:** Motion dataset description.

| Motion: | Row | Run | Stretch | FH | BH | B&F | Serve1 | Serve2 | Serve3 | Serve4 |
|---|---|---|---|---|---|---|---|---|---|---|
| Length (min) | 7 | 13 | 9 | 0.5 | 0.5 | 1.33 | 1.5 | 1.5 | 1.25 | 1.25 |

Our dataset is illustrated in Figure 55 and 56. Figure 55 shows the selected examples of our periodic to quasi-periodic motion collection, including rowing, running, stretching, the practice of forehand (FH) and backhand (BH) in the racquetball play, and the interlaced forehand and backhand (B&F). They all have short time-scale periods. Various elements of quasi-periodicity exists in this dataset. For example, the person may change the speed, answering phone call or drinking watering while running, change her position during stretching exercise. The forehand and backhand practice are quasi-periodic since the player is not a professional and may hit the ball with different speed and gestures.

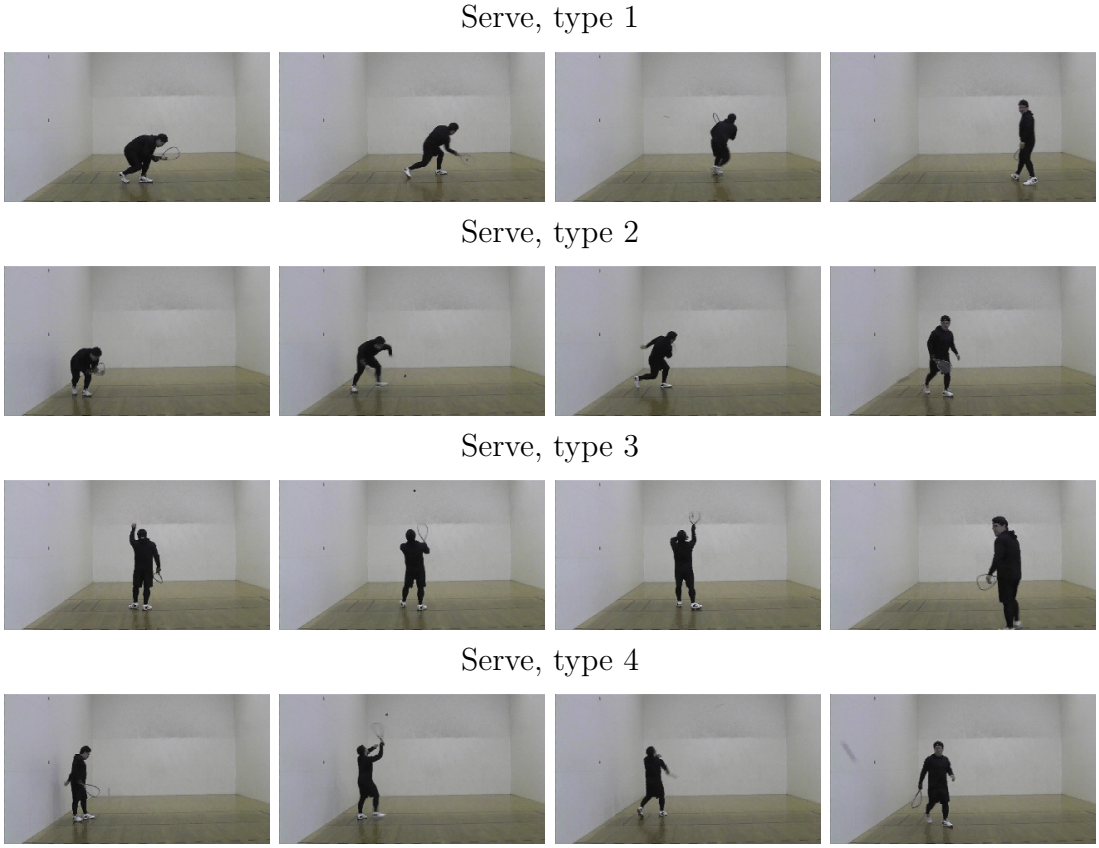


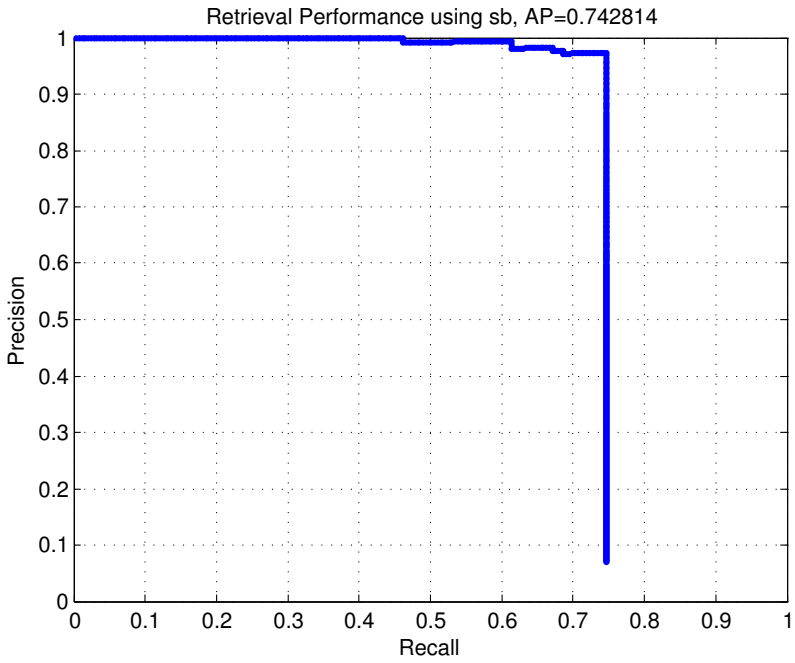

**Figure 56:** Selected quasi-periodic motions from Racquetball practice.

Figure 56 shows the quasi-periodic activities of the practice of racquetball serves. Each period consists of a series of short time-scale actions. There are four types of serves, each hitting towards a different corner on the wall. Table 12 lists the lengths of each action category in the dataset.

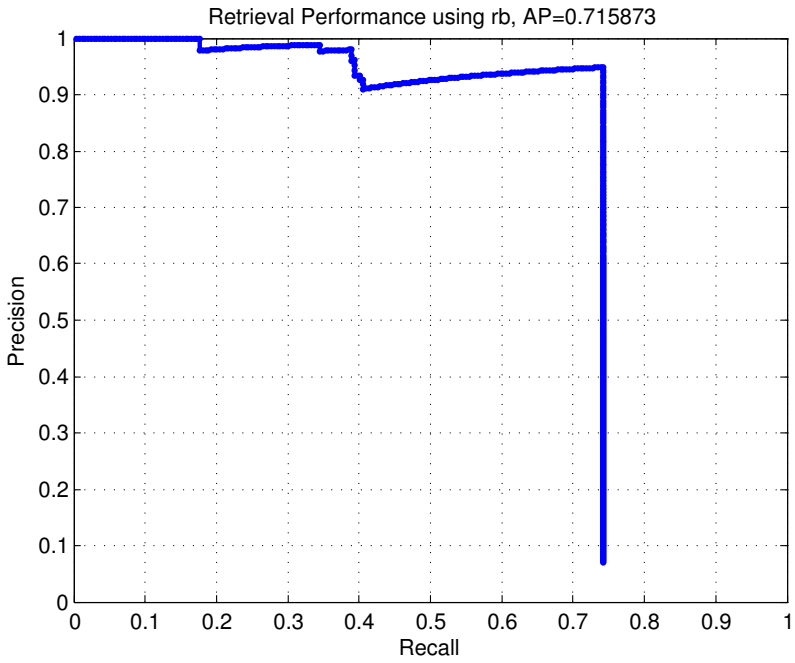
Our non-periodic motions consist of the *nongame* videos from the ChildPlay dataset and the *noninteraction* videos from our home movie collection. In total we have 226 (quasi-)periodic motion segments and 2387 non-periodic video segments.

### 6.3.2 Overview of the retrieval performance

Our quasi-periodic pattern mining method demonstrates much superior retrieval performance over self-similarity based approach for all categories of the motions. Overall, the self-similarity approach has little tolerance to quasi-periodicity, reflected by the



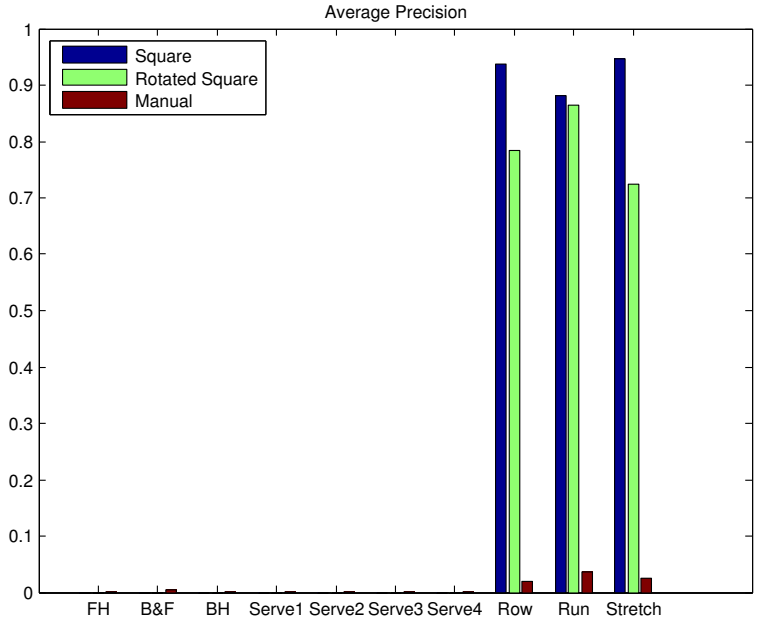

(a) Match with square lattice, $AP = 0.742814$



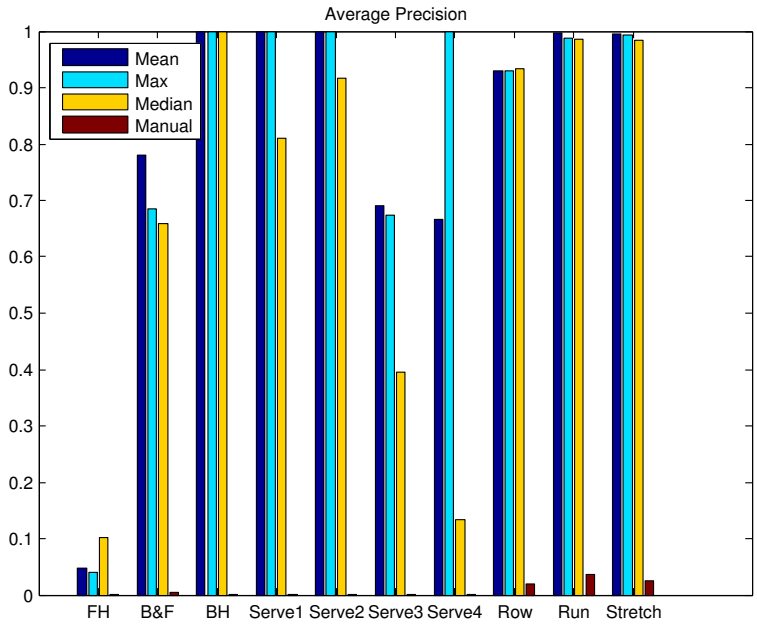

(b) Match with rotated square lattice, $AP = 0.715873$

**Figure 57:** Precision-recall curves based on self-similarity measurement.

fact that it never has a 100% recall rate for any motion category. We give detailed performance measurements in terms of precision-recall curves and average precisions




(a) Retrieval based on Self-similarity

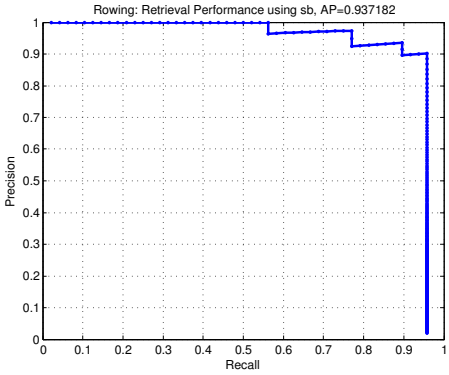



(b) Retrieval based on quasi-periodic pattern mining

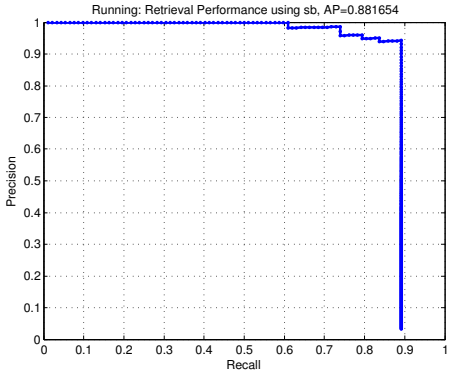
**Figure 58:** Average Precision for each motion category.

for both methods.

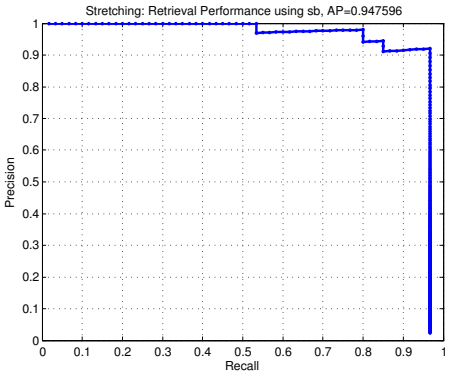
**Retrieval based on self-similarity measurement** Figure 57 shows the retrieval




(a) Rowing




(b) Running




(c) Stretching

**Figure 59:** Retrieval performance for each motion category using Cutler's method.

performance using Cutler's method with both the square lattice and the rotated square lattice structures. We make two observations on the retrieval results. First, there are periodic motions that can't be retrieved at all for both lattice structures, which means that Cutler's method tends to have high false negatives. Second, the precision is high (above 90%) before it reaches the breaking down point, where no more true positives can be retrieved. This shows that this method is good at rejecting false predictions of periodicity.

Figure 58(a) shows the AP for each category of motion using Cutler's method. "Manual" is the AP value for the case of manual search. None of the racquetball practice can be retrieved due to its "quasi-periodicity". Figure 59(a), 59(b) and 59(c) are the precision-recall curves based on matching the square lattice for rowing, running

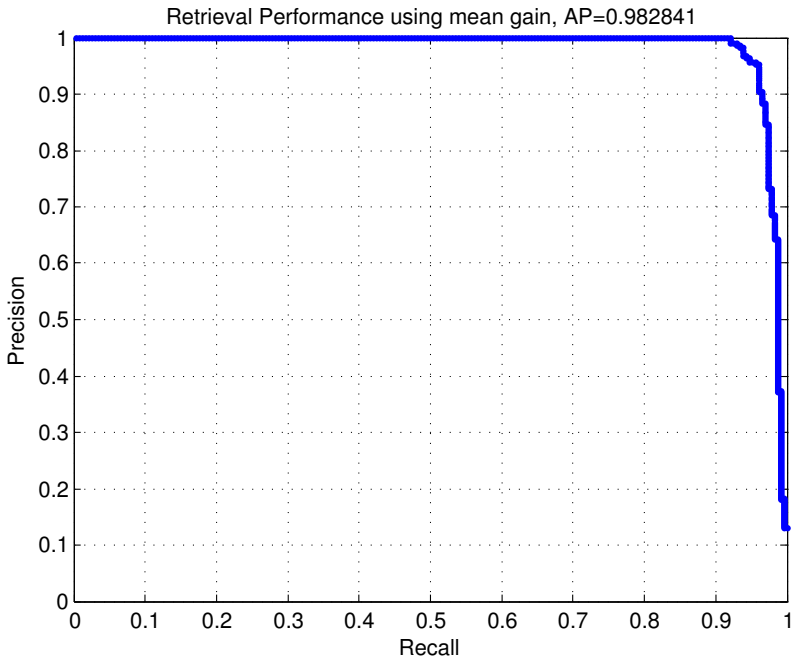



**Figure 60:** Retrieval performance of Quasi-Periodic Pattern Mining using $mean(G)$.

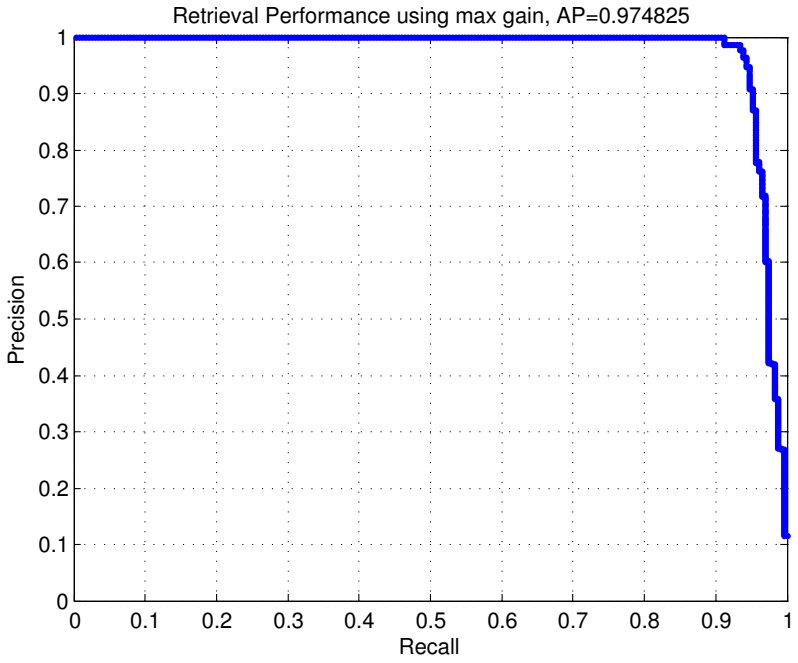



**Figure 61:** Retrieval performance of Quasi-Periodic Pattern Mining using $max(G)$.

and stretching exercise respectively. Again, this method has high precision, but tend to make false rejections, even for highly periodic motions such as bowing, running,

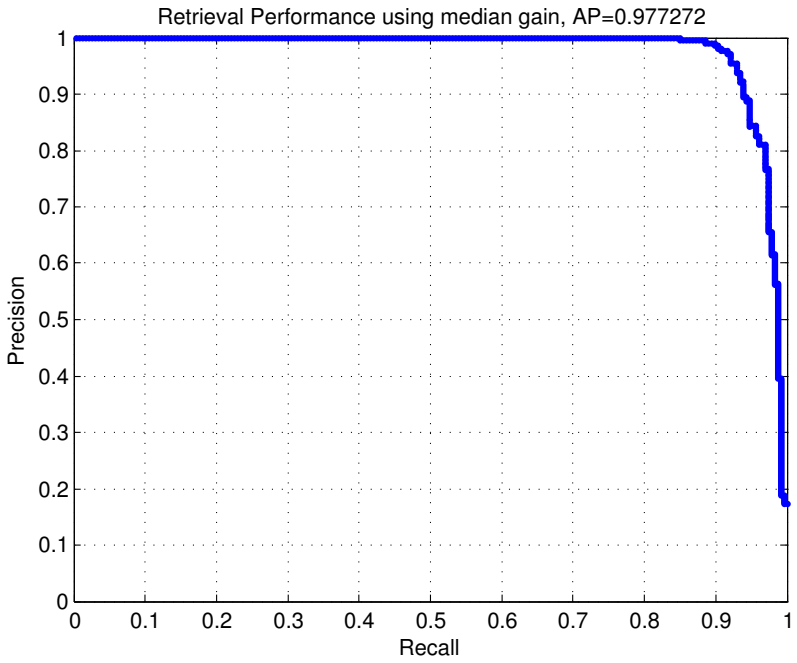
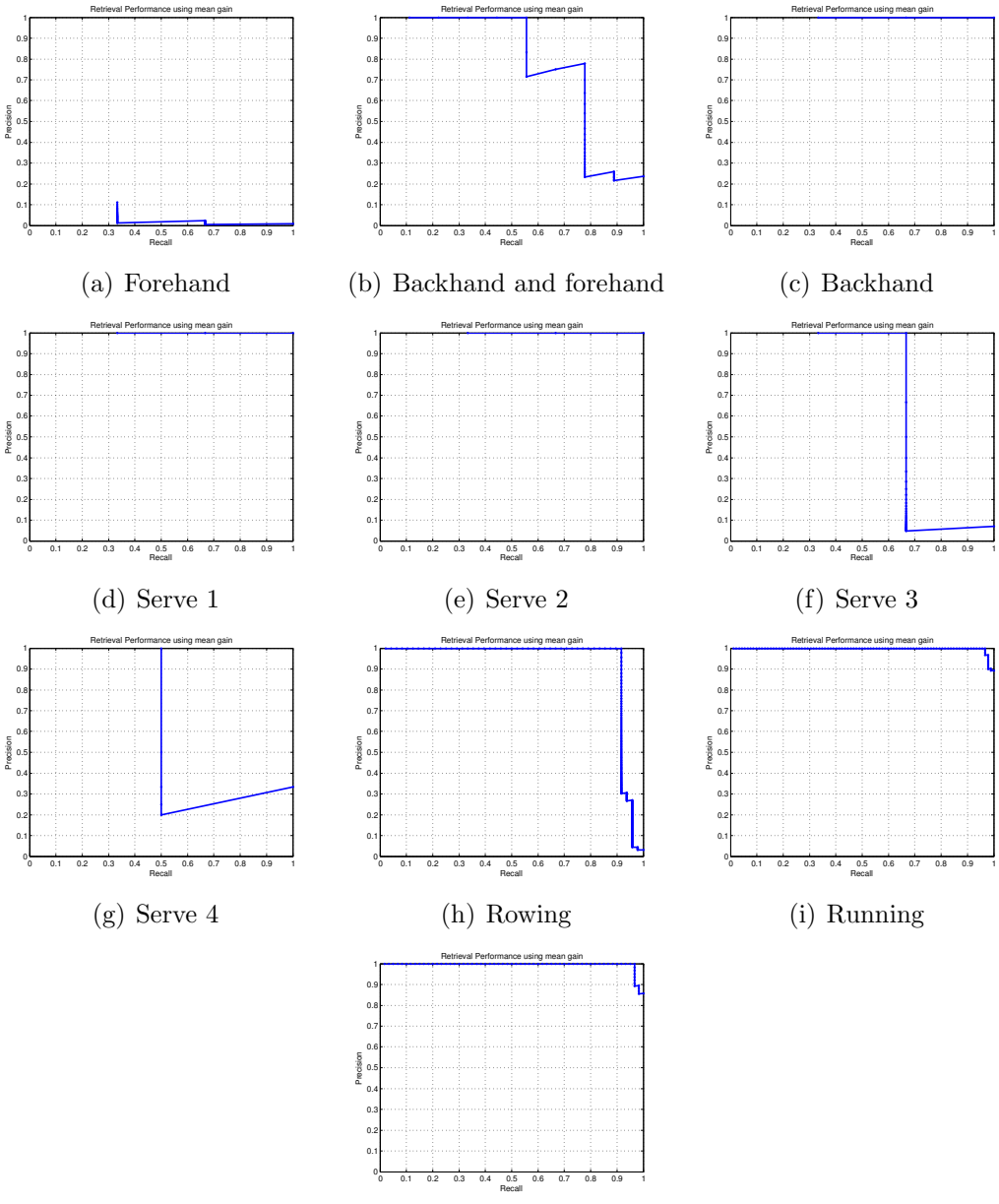



**Figure 62:** Retrieval performance of Quasi-Periodic Pattern Mining using $median(G)$.

and stretching.

**Retrieval based on quasi-periodic pattern mining** Figure 60, 61 and 62 show the precision-recall curves measured according to $mean(G)$, $max(G)$ and $median(G)$ respectively. They all achieve a very good retrieval performance. The retrieval ranked with $mean(G)$ has a precision of 100% when the recall ranges from 0 to 92%. For all the three measurements, our method can retrieve all the relevant items.

Figure 58(b) shows the average precision for each motion category. The precision-recall curves based on $mean(G)$ for each category are shown in Figure 63. We observe that our method retrieves all the instances of "Backhand", "Serve 1" and "Serve 2" without introducing any false negatives. In contrast, Cutler's method fails to retrieve any of them, due to the non-constant period and the possible changing appearances in different periods. The retrieval performance for "Running" and "Stretching" is better than that of Cutler's method. For "Running", the precision is 100% when the recall is up to 96.74%, while Cutler's method has a precision around 94% when the




(a) Forehand (b) Backhand and forehand (c) Backhand

(d) Serve 1 (e) Serve 2 (f) Serve 3

(g) Serve 4 (h) Rowing (i) Running

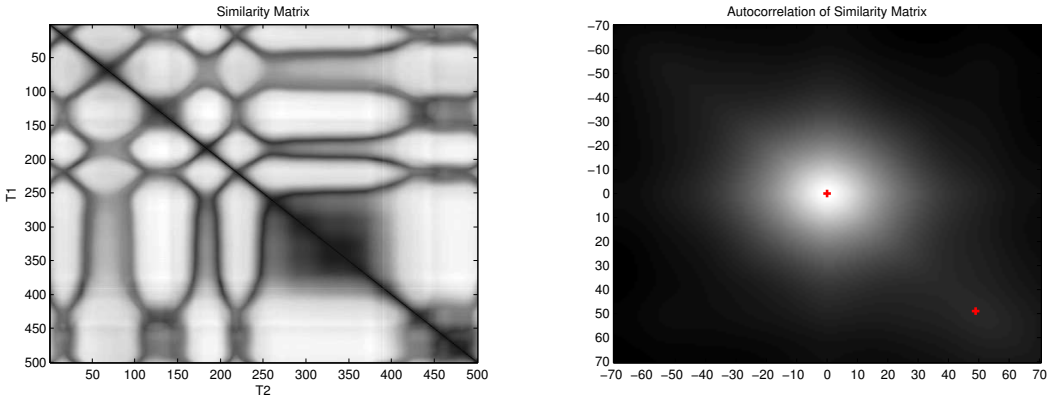
(j) Stretching

**Figure 63:** Retrieval performance for each motion category using Quasi-Periodic Pattern Mining.

recall reaches its maximum of 89%. For "Stretching", Cutler's method has a precision of 92.06% when recall reaches its maximum 96.67%; our method still has precision 100% when the recall is 96.67%. For "Rowing", Cutler's method has a slightly higher AP (0.937182) than that of our method ($AP = 0.930248$).

### 6.3.3 False negatives of self-similarity based approach

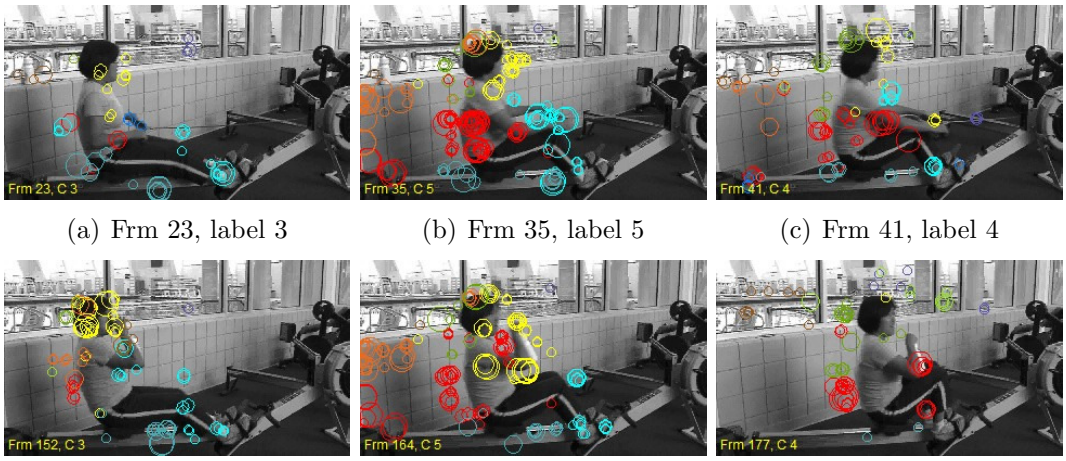
**Figure 64:** Selected frames of a false negative from a rowing motion sequence.




(a) Similarity plot

(b) Autocorrelation matrix

**Figure 65:** Similarity and autocorrelation plots for a bowing sequence shown in Figure 64.




(a) Frm 23, label 3

(b) Frm 35, label 5

(c) Frm 41, label 4

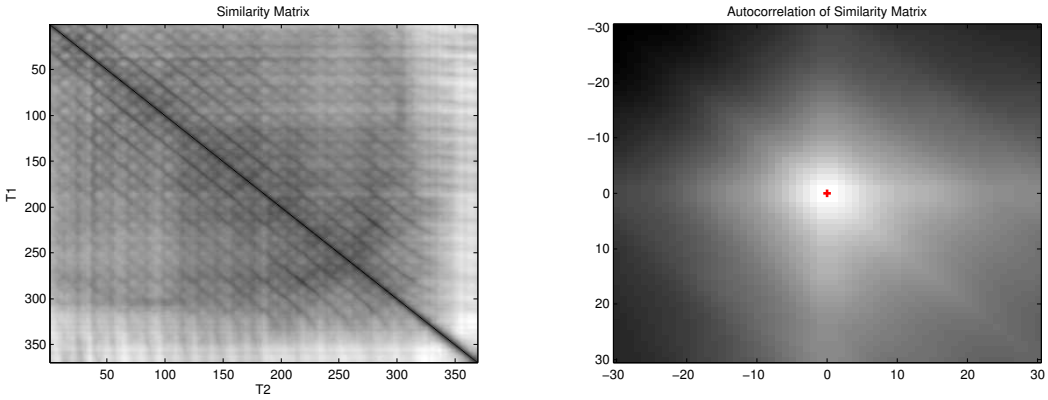
(d) Frm 152, label 3

(e) Frm 164, label 5

(f) Frm 177, label 4

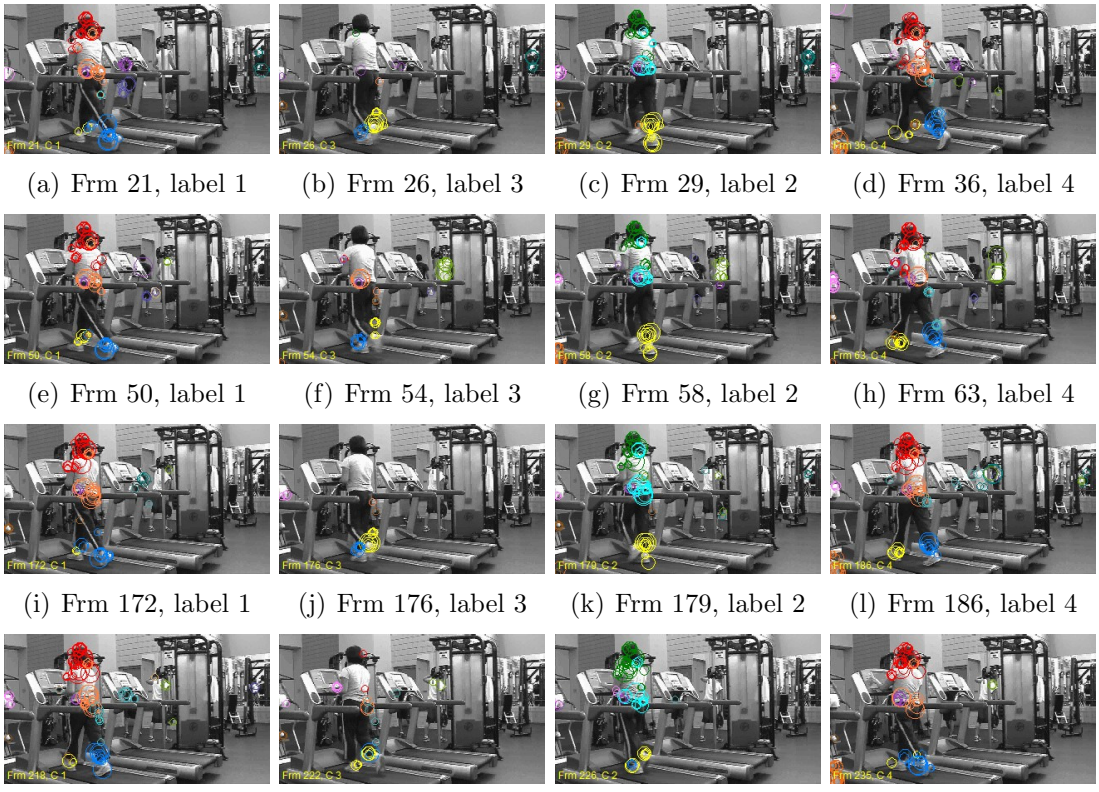
**Figure 66:** Pattern 3-5-4 extracted from the sequence shown in Figure 64.

We give detailed analysis on the false negatives of Cutler's method at the point where no more relevant items can be retrieved.




(a) Similarity plot (b) Autocorrelation matrix

**Figure 67:** Similarity and autocorrelation plots for a running sequence.



(a) Frm 21, label 1 (b) Frm 26, label 3 (c) Frm 29, label 2 (d) Frm 36, label 4

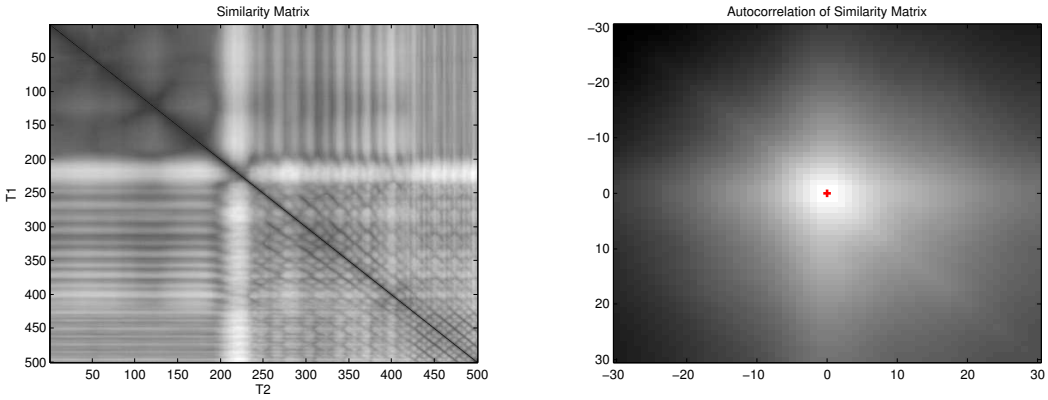
(e) Frm 50, label 1 (f) Frm 54, label 3 (g) Frm 58, label 2 (h) Frm 63, label 4

(i) Frm 172, label 1 (j) Frm 176, label 3 (k) Frm 179, label 2 (l) Frm 186, label 4

(m) Frm 218, label 1 (n) Frm 222, label 3 (o) Frm 226, label 2 (p) Frm 235, label 4

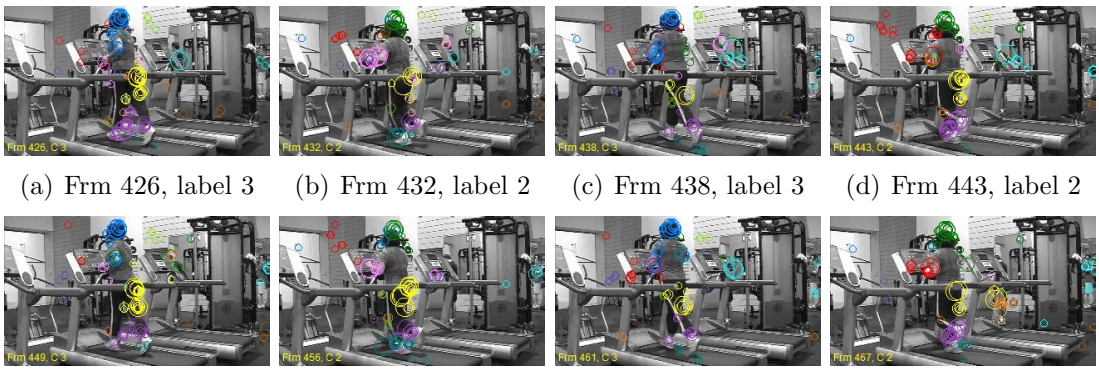
**Figure 68:** Four occurrences of Pattern 1-3-2-4 extracted from a running sequence with the similarity plot in Figure 67.

**Analysis of rowing** Figure 65 shows the similarity plot and the autocorrelation matrix for a false negative in the rowing exercise (Figure 64. The person in this sequence starts with a slowing-down rowing motion (Figure 64(a)), then stops to take




(a) Similarity plot (b) Autocorrelation matrix

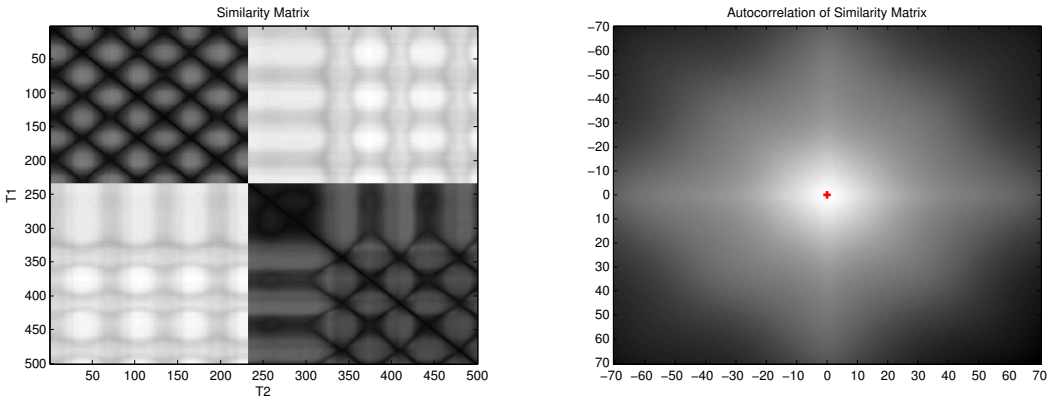
**Figure 69:** Similarity and autocorrelation plots for a running sequence.



(a) Frm 426, label 3 (b) Frm 432, label 2 (c) Frm 438, label 3 (d) Frm 443, label 2

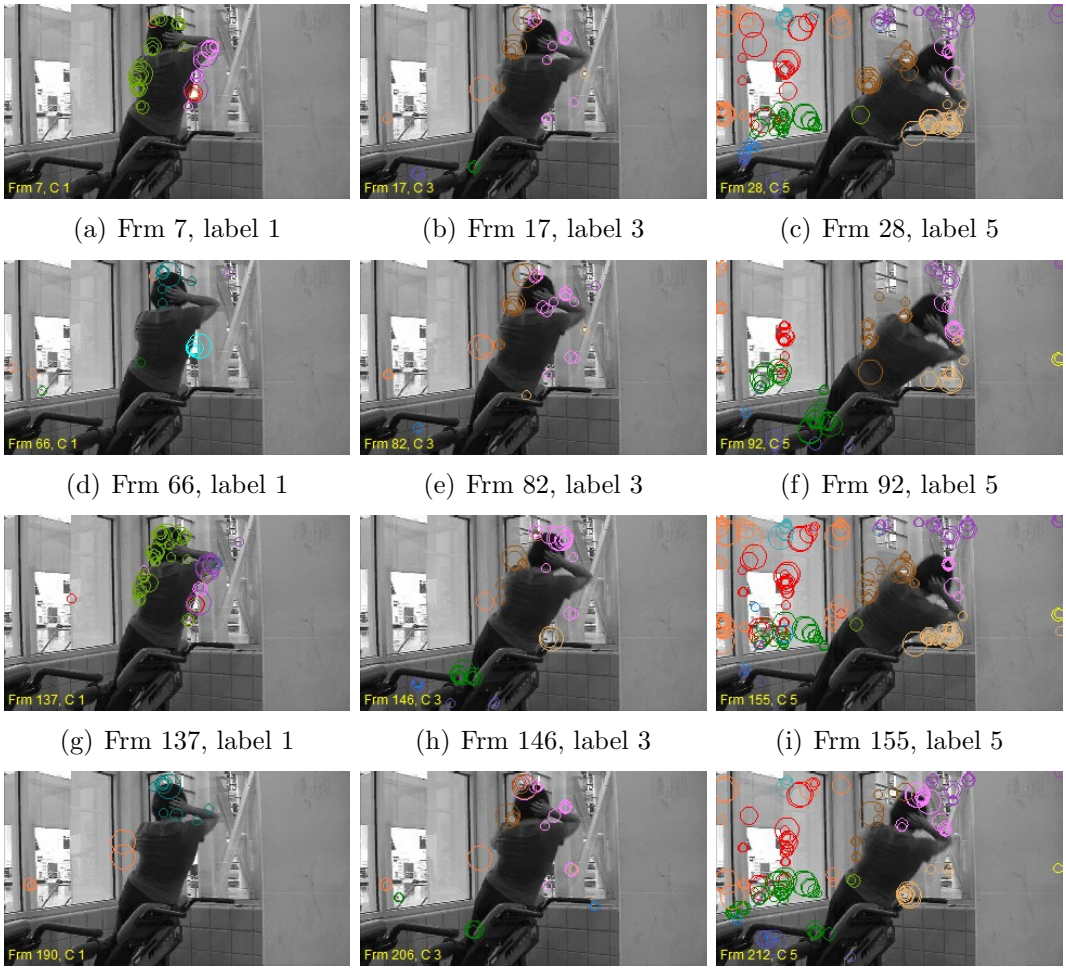
(e) Frm 449, label 3 (f) Frm 456, label 2 (g) Frm 461, label 3 (h) Frm 467, label 2

**Figure 70:** Pattern 3-2 extracted from a running sequence with the similarity plot in Figure 69.




(a) Similarity plot (b) Autocorrelation matrix

**Figure 71:** Similarity and autocorrelation plots for a stretching sequence.




(a) Frm 7, label 1 (b) Frm 17, label 3 (c) Frm 28, label 5

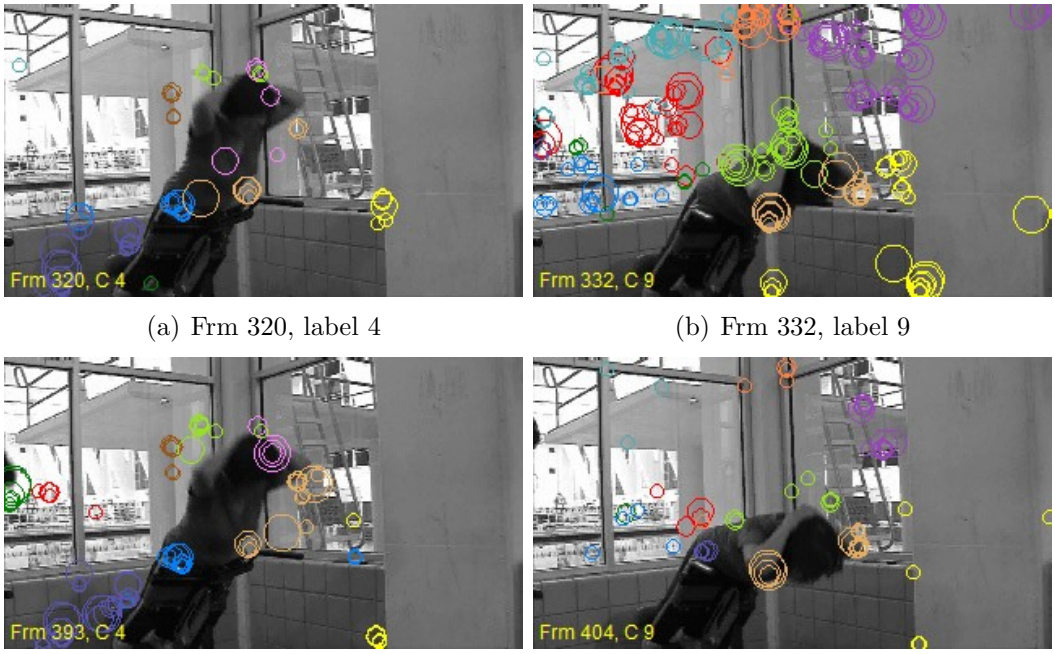
(d) Frm 66, label 1 (e) Frm 82, label 3 (f) Frm 92, label 5

(g) Frm 137, label 1 (h) Frm 146, label 3 (i) Frm 155, label 5

(j) Frm 190, label 1 (k) Frm 206, label 3 (l) Frm 212, label 5

**Figure 72:** Pattern 1-3-5 extracted from a stretching sequence with the similarity plot in Figure 71.

her bottle (Figure 64(b)), moving back and forth while drinking the water, and put back the bottle, then prepare to row again (Figure 64(d)). The similarity plot shows that the period is highly non-constant, which is a challenge to STFT and the matching of the structure of local peaks in $A$. The quasi-periodicity in this sequence lies in the three aspects: 1) the varying period $T$. Since the person moves subconsciously back and forth while drinking the water, the period is not constant; 2) the varying poses at different cycles. In the beginning the person is still in the exercise state, pulling the handles while rowing. After taking the bottle, she moves without pulling the handles; 3) extraneous actions of taking the bottle and putting it back.




(a) Frm 320, label 4 (b) Frm 332, label 9

(c) Frm 393, label 4 (d) Frm 404, label 9

**Figure 73:** Pattern 4-9 extracted from a stretching sequence with the similarity plot in Figure 71.

In contract, our QP method can detect patterns of different periodic motions from this sequence. Figure 66 shows the two occurrences of pattern 3-5-4 from the same sequence. Figure 66(a) to 66(c) captures a rowing exercise, while Figure 66(d) to 66(f) captures the motion of rowing while drinking. It successfully filters out the extraneous action of taking the bottle.

**Analysis of running** We give two examples of false negatives using Cutler's method.

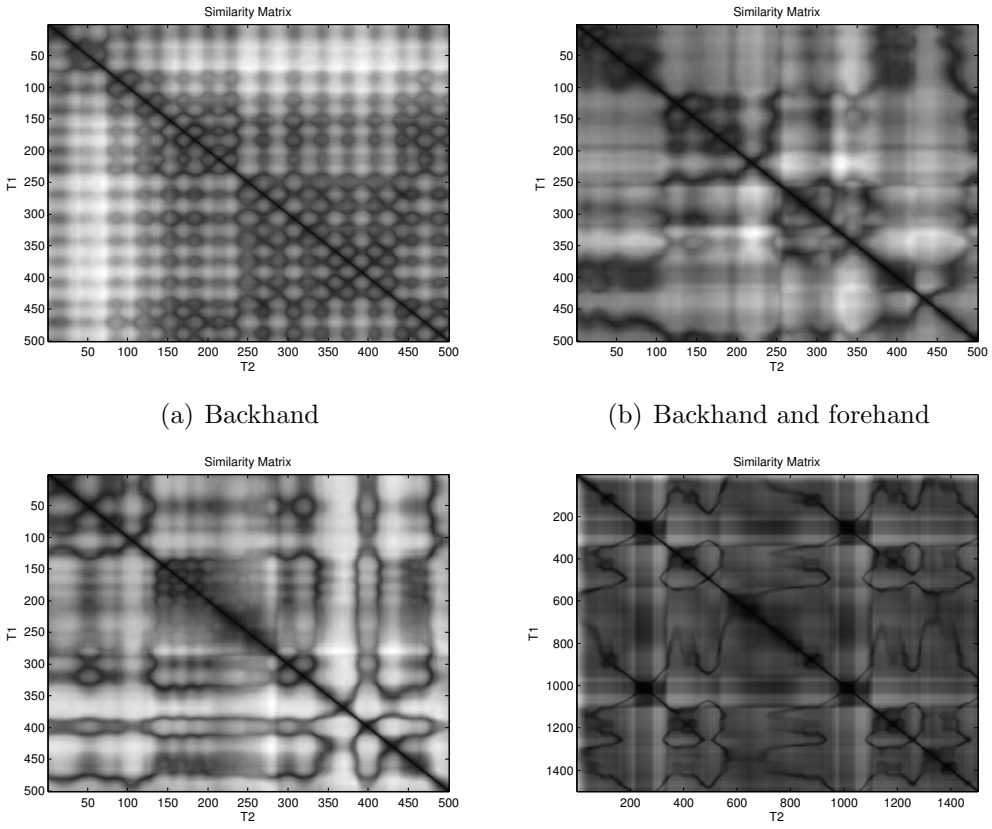
Figure 67(a) shows the similarity plot of a running sequence. We can see several dark lines parallel to the diagonal, which indicates the existence of periodic motions. The similarity plot is also noisy due to other people's random motions in the background. In the end of the sequence, due to a person walking towards the camera, the difference between two frames goes up very quickly. Cutler and Davis [18] points out that one can carefully choose the window size of STFT to estimate the local period $d$, then construct the autocorrelation matrix $A$ by correlating the proper part of the similarity matrix.

Our method, on the other hand, does not require any knowledge of the periodicity and where it occurs in the sequence. Figure 68 shows the four occurrences of pattern 1-3-2-4 that corresponds to the highly periodic running actions in that sequence. Note that the perturbation from the extraneous movements in the background does not affect the periodic pattern mining, since the foreground motion dominates the generation of interest points.

Another example of false negative is shown in Figure 69. The person takes a break and stands on the treadmill in the beginning, then gradually changes from walking to running. This trend is clearly seen in the similarity plot 69(a). Our method extracts the pattern 3-2 that corresponds to the running motion, shown in Figure 70, where label 3 corresponds to the moments when the two legs moving apart, and label 2 corresponds to the moments when the two legs moving towards each other. This example demonstrates the superiority of our method over Cutler's method at detecting periodicity over a sequence that contains varying periods.

**Analysis of stretching** Our video collection of stretching exercise contains highly periodic motions and Cutler's method successfully retrieves most of them with high precision. Figure 71 shows a similarity plot and $A$ from a stretching sequence that is falsely rejected by Cutler's method. This sequence contains two different stretching exercises, each of which is highly periodic. Again, if one knows the beginning and ending of each periodic motion and apply STFT to the columns of $S$, one can estimate the period of each motion. Or using that knowledge to calculate the autocorrelation matrix of the proper submatrix of $S$.

In contrast, our method can find all the periodic motions (of different content) from the sequence without any such knowledge. Figure 72 and Figure 73 show the two patterns that correspond to the two different stretching exercises in the sequence.




(a) Backhand

(b) Backhand and forehand

(c) Forehand

(d) Serve 1

**Figure 74:** Similarity plots from the racquetball practice.

### 6.3.4 Quasi-periodic patterns mined from the racquetball sequences

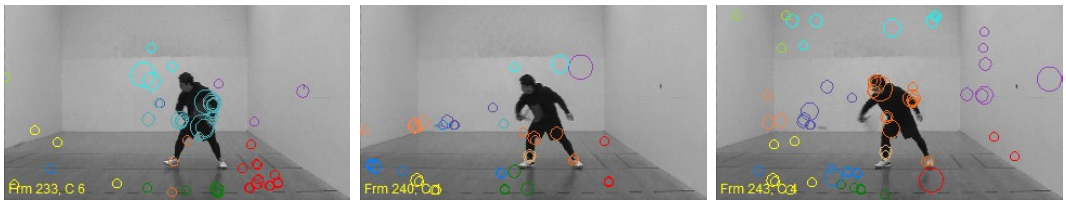
Self-similarity method fails to retrieve any quasi-periodic motions from the racquetball practice. Figure 74 shows the similarity plots from the practice of backhand, forehand, interlaced backhand and forehand, and type 1 serve. All the sequences demonstrate certain degree of quasi-periodicity. Our method, on the contrary, finds all the instances of backhand practice, serve 1 and 2 without any false positives, and is able to produce precision-recall curves with full-range of recall rate for all the racquetball practice sequences.

**Backhand practice** Figure 75 shows 6 of 8 occurrences of the pattern 6-1-4 from the backhand practice. This sequence is highly periodic, as the player tries to hit the ball to the same spot on the wall, and hits gently to stay at the same spot. The

(a) Frm 37, label 6 (b) Frm 41, label 1 (c) Frm 47, label 4

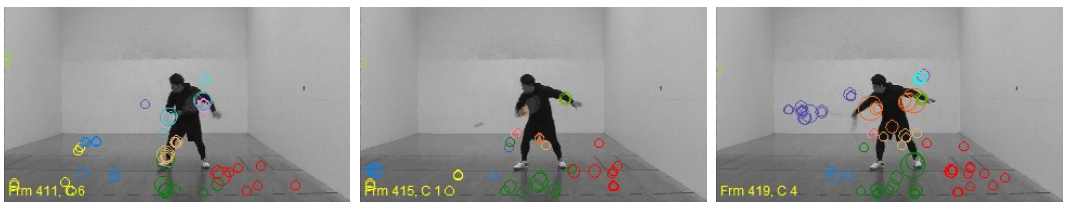
(d) Frm 107, label 6 (e) Frm 109, label 1 (f) Frm 114, label 4

(g) Frm 139, label 6 (h) Frm 142, label 1 (i) Frm 148, label 4



(j) Frm 233, label 6 (k) Frm 240, label 1 (l) Frm 243, label 4

(m) Frm 311, label 6 (n) Frm 315, label 1 (o) Frm 320, label 4



(p) Frm 411, label 6 (q) Frm 415, label 1 (r) Frm 419, label 4

**Figure 75:** Pattern 6-1-4 from a backhand practice sequence.

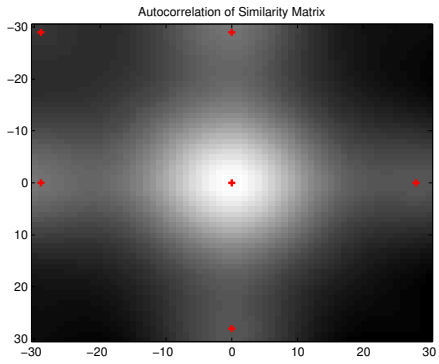
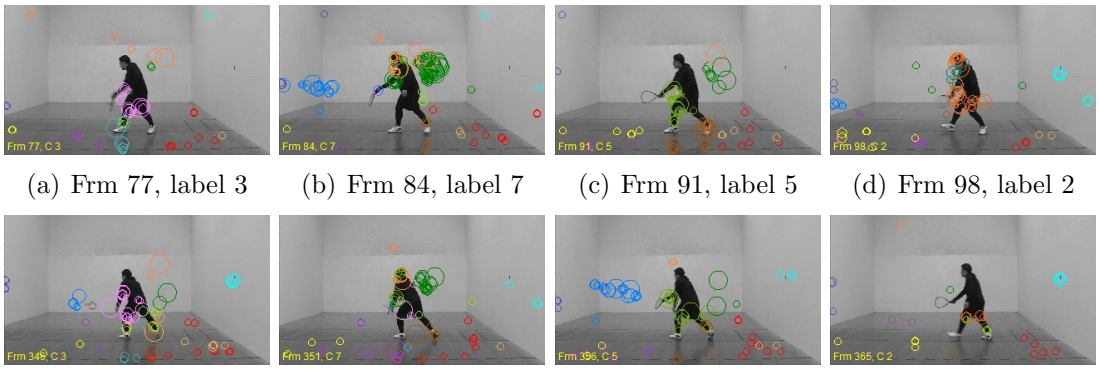


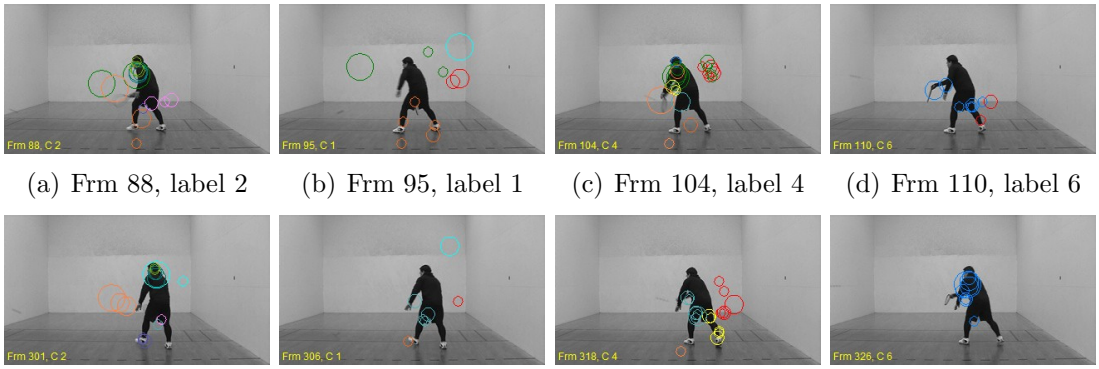

**Figure 76:** Autocorrelation matrix of the similarity plot shown in Figure 74(a). Local peaks are indicated by red + symbols.




(a) Frm 77, label 3 (b) Frm 84, label 7 (c) Frm 91, label 5 (d) Frm 98, label 2

(e) Frm 346, label 3 (f) Frm 351, label 7 (g) Frm 356, label 5 (h) Frm 365, label 2

**Figure 77:** Pattern 3-7-5-2 from a sequence of interlaced backhand and forehand practice.

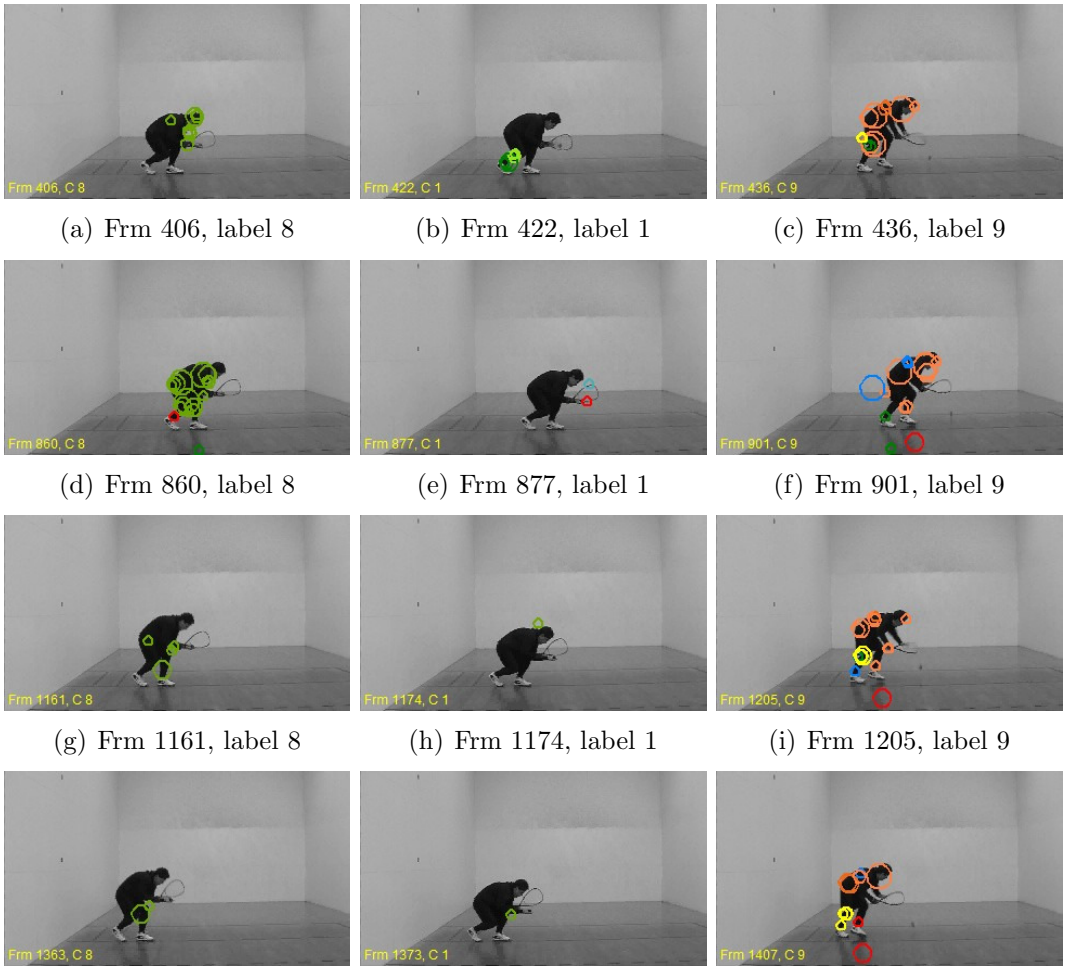



(a) Frm 88, label 2 (b) Frm 95, label 1 (c) Frm 104, label 4 (d) Frm 110, label 6

(e) Frm 301, label 2 (f) Frm 306, label 1 (g) Frm 318, label 4 (h) Frm 326, label 6

**Figure 78:** Pattern 2-1-4-6 from a sequence of forehand practice.

similarity plot in Figure 74(a) validates the existence of periodicity. It also shows that the motions in the beginning of this sequence is much different from other frames,



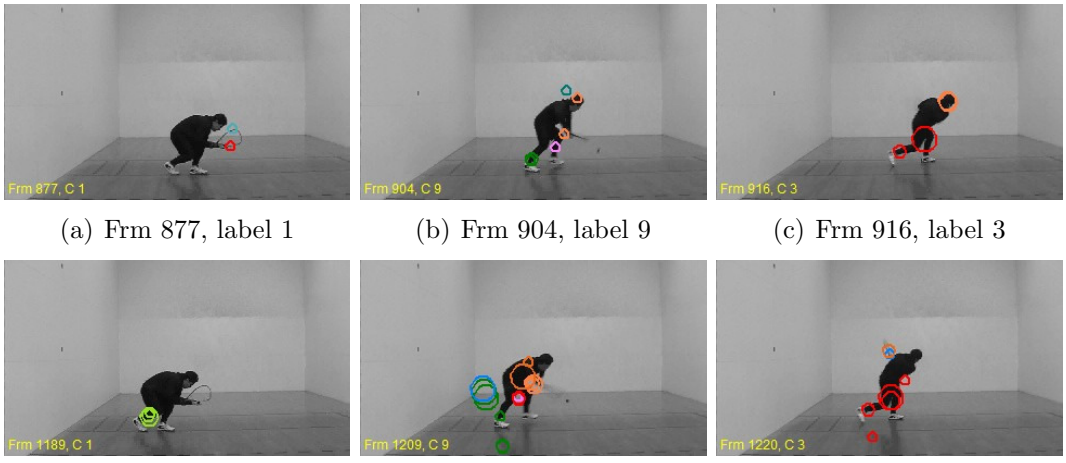

(a) Frm 406, label 8 (b) Frm 422, label 1 (c) Frm 436, label 9

(d) Frm 860, label 8 (e) Frm 877, label 1 (f) Frm 901, label 9

(g) Frm 1161, label 8 (h) Frm 1174, label 1 (i) Frm 1205, label 9

(j) Frm 1363, label 8 (k) Frm 1373, label 1 (l) Frm 1407, label 9

**Figure 79:** Pattern 8-1-9 from a sequence of serve practice.




(a) Frm 877, label 1 (b) Frm 904, label 9 (c) Frm 916, label 3

(d) Frm 1189, label 1 (e) Frm 1209, label 9 (f) Frm 1220, label 3

**Figure 80:** Pattern 1-9-3 from a sequence of serve practice.

as the player is preparing for the practice. The differences affect the autocorrelation matrix, as shown in Figure 76, which presumably should have peaks with much closer intervals.

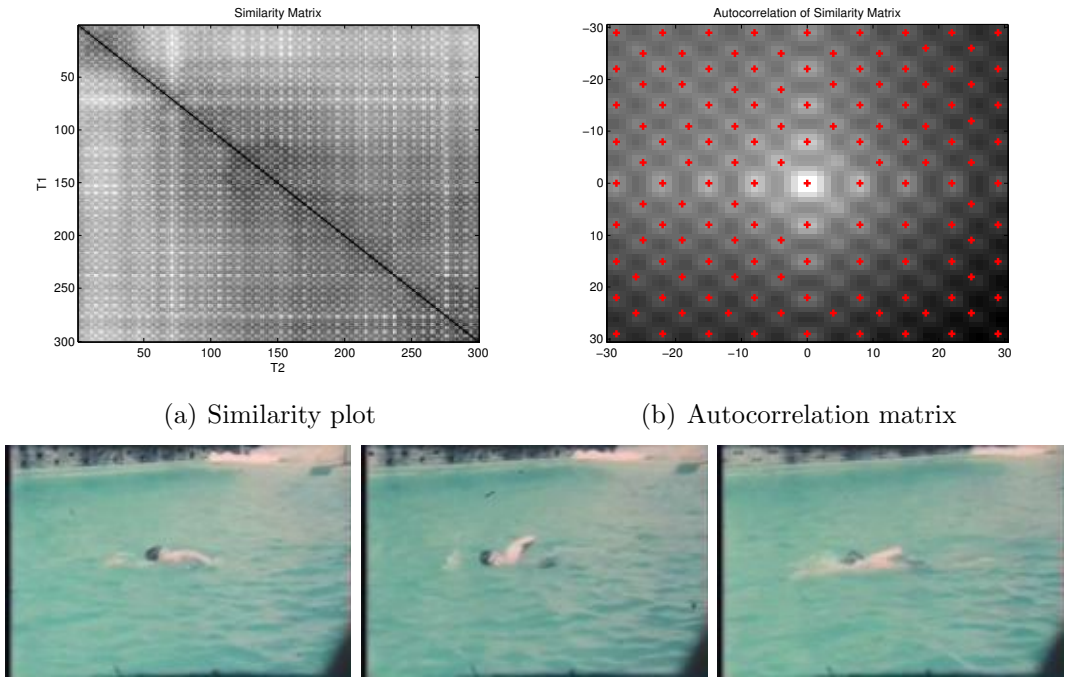
**Interlaced backhand and forehand practice** In this sequence, the player tries to hit the ball to the same spot on the wall, using forehand and backhand in turn. Since the player is not a professional, the ball is not controlled precisely and his poses and speed of hitting-the-ball vary a lot. Occasionally, the position of the player and the position of the ball make it difficult to change from forehand to backhand or vice versa, the player may do several backhand or forehand in a row before switching. This quasi-periodicity is demonstrated by the similarity plot shown in Figure 74(b). Again our method successfully finds the patterns that capture the recurring practice. Figure 77 shows the pattern 3-7-5-2 that captures the forehand in this sequence.

**Forehand practice** The quasi-periodicity (Figure 74(c)) in the forehand practice is from the varying period and poses, as the player moves much more than he does the backhand practice. Our method is able to retrieve the quasi-periodic motions, while the precision rate is relatively low compared to other racquetball practice. Figure 78 shows the pattern 2-1-4-6 extracted from a forehand practice.

**Serve practice** A serve is a sequence of short-term actions, and it typically involves four steps: prepare, hit, pick-up and ball, get-ready again (Figure 56). The quasi-periodicity in the practice of serves (Figure 74(d)) comes from the variations of the period, and the poses and duration at each stage of a serve. There is no extraneous actions in the background. Our method is very suitable for detecting quasi-periodic patterns for such sequences that involves recurrences of a series of actions. Whether the quasi-periodic pattern mining can find the pattern that captures every stage of a serve depends on the proper assignment of the visual words and the detections of critical interest points at every occurrence. Most of the time, the patterns extracted by our method capture the various subset of the stages. For example, Figure 79 and 80

show the two patterns that are found from the same sequence, each parsing a subset of a complete serve. Pattern 8-1-9 captures the sequence of bending-down, holding-still, and swinging. Pattern 1-9-3 captures the sequence of holding-still, swinging and the finish of a swing.

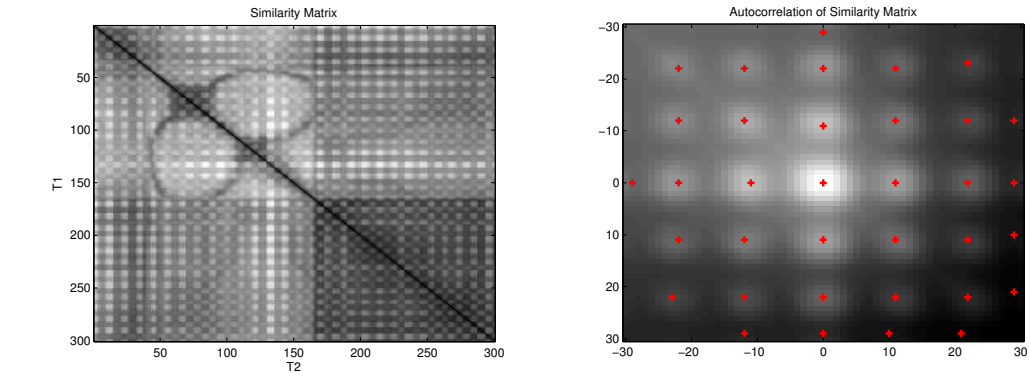
### 6.3.5 False positives of both methods




(a) Similarity plot (b) Autocorrelation matrix

**Figure 81:** A retrieved instance of swimming from our home movie collection.

For Cutler's method, there are only 5 false positives based on the square lattice matching when the retrieval reaches the breaking-down point. 4 out of the 5 are from a swimming sequence, which are actually quasi-periodic motions, not a recurring social interaction though (see Section 4.3.3). Figure 81 shows the similarity plot and the autocorrelation matrix with detected local peaks. Figure 82 shows the 5th false positive detected by Cutler's method. Recorded in the dark, this video contains a group of children hanging around a Christmas tree to find their presents. The periodicity of image pixels is probably caused by the high contrast between the moving foreground and background, and the fact that the videographer moves the camera




(a) Similarity plot (b) Autocorrelation matrix

**Figure 82:** A false positive retrieved from our home movie collection.

smoothly and keeps the moving children in the frame center.

For our approach, at the operating point with $recall = 96.46\%$ and $precision = 90.08\%$ (Figure 60), the sequences of swimming were also detected as false positives. The sequence shown in Figure 82 was correctly rejected due to its lack of quasi-periodic patterns with sufficient scores. On the other hand, our method retrieved other false positives from home movies that do exhibit repeating motions with varying periods, which can be found in Figure 16.

## *6.4 Discussions*

The notion of quasi-periodicity extends the concept of periodic motion to a much broader range of actions that are repetitive. Its three elements allow much bigger within-class variations of poses and periods, and the random insertion or deletion of actions. It also includes the case where each period lasts more than several minutes and consists of a series of short time-scale actions.

Detection and analysis of such quasi-periodic motions in unstructured videos is much more challenging than periodic motion detection. Previous works on extracting

recurring actions only process accurate action signal, such as the motion capture data [100], instead of the raw pixels from videos directly. Many conventional periodic motion analysis methods extract information from videos directly, but the actions of interest are repetitions of short time-scale actions such as walking and running [18, 41, 66]. These methods often needs to estimate the constant period, and involve a lot of parameter selection.

In contrast, our quasi-periodic pattern mining approach operates on the video pixels directly, and can extract quasi-periodic patterns of any length without any prior knowledge about where and how the quasi-periodicity will occur. The current performance depends on the correct assignment of frame labels. However, our algorithm provides a general framework for recurring time series (of any length) extraction. One can keep improving the video tokenization and still apply our quasi-periodic pattern mining method to find quasi-periodic motions. In the future, we are interested in finding quasi-periodic actions in videos without tokenizing the videos. For example, we can find similar spatio-temporal structure sequences across the video. This framework is also applicable to detect recurring speeches [63] or texts or multi-media signals.

# CHAPTER VII

# CONCLUSIONS AND FUTURE WORKS

This thesis was motivated by the strong desire for efficient video filtering from psychologists in their work on early detection of autism. Our approach for social game analysis in unstructured videos is applicable to video-based social interaction analysis for behavior science in general. Parent-child social games are characterized by repetitions of dyadic turn-taking interactions, with a permissable range of variations. My thesis statement is:

> Repetition of turn-taking in dyadic social games, as revealed by space-time sequential motion structures, supports effective retrieval of these games from unstructured real-world videos.

The thesis work has made four contributions:

1. A new problem to computer vision — analysis of social interactions in unstructured videos. We have published the ChildPlay video dataset and the annotations to the research community to support and encourage further research on human behavior analysis in realistic contexts.

2. A computational model of social games as quasi-periodic events in a time series, and an unsupervised algorithm that mines the quasi-periodic patterns in videos. This method represents a substantial generalization of conventional periodic motion analysis.

3. A effective categorization of YouTube videos of social games based on bag-of-words model and quasi-periodic pattern mining that selects characteristic visual words automatically.

4. A advocate for more research in order to understand human behaviors, and make fundamental influence to the current diagnosis and treatment of behavioral and developmental disorders.

The efficacy of our approach leverages on two facts. First, the crucial motions in social games, which often involve sudden change of movement directions and velocities, generate corners in the 3D space-time volume of videos. Such corners can be detected with certain reliability in real-world videos, and make it possible to represent the behaviors with sparse visual words. Second, the repetitive temporal structure of the social games over a long time scale is a powerful cue for unsupervised activity pattern discovery. A number of future works can be done to improve the retrieval performance, and to provide intelligent support for the understanding of the dyadic interactions.

## *7.1 Modeling the Turn-taking Interactions*

The interaction generates the causal relationship between the behaviors from the parent and the child. Prabhakar *et al.* have shown that the retrieval performance can be improved by analyzing the temporal causalities among the multi-points process of the visual words [67]. Modeling the spatial and temporal causal relationship is the next work to improve the performance of social game retrieval, as the interactions take place not only in time, but also in space.

Audio signal provides another informative channel for social game retrieval and synchrony analysis. Most parents play games with strong vocal signals (*e.g.*, "peek-a-boo!", "I will get you"), facial expressions, and gestures to engage their children. Such audio signals also exhibit a temporal structure, and should be combined with the visual signals for a better characterization of the interactions.

Visual and audio signals provide contextual cues for each other, so are the objects in the scene. For example, a ball flying in the air or rolled on the floor back and forth

increases the probability of a ball game is being played. The parents' actions may be less variant than the children's when they are trying to initiate the game, and we can incorporate supervised detection or prior knowledge to detect the actions from the parents and predict the responses from the children. The furniture layout (when analyzing home videos) also has an impact on the parent or the child will move. For example, one may hide behind a sofa in a peek-a-boo game. In short, the scene, the available toys and objects in the scene, and the actions from both the parent and the child may affect each other, and the ability to capture interactions among these factors will greatly enhance our ability to explain and understand the behaviors in natural environments.

## *7.2 Automatic Quantification of Interactions*

Turn-taking analysis shall prepare us well for the automatic quantification of the interactions, such as how well the child is engaged, does the child initiate the games often, how good is he at expressing his social request, *etc.* The measurement of the interactions can be conducted in different environments, ranging from structured, clinic-based interaction, to natural interactions at home, daycare, or schools. With the varying ability to capture the gaze shifting, or facial expressions, or body gestures under different environments, we can give characterizations of the children's interactions with different levels of details.

# APPENDIX A

# SUFFIX TREE

Suffix tree is a data structure that was first proposed by Weiner [88] for presenting all the suffixes of a given string in a way that allows fast implementation of many string operations. In this chapter, we will review the space-efficient method proposed by McCreight [50], with a focus on the intuitive explanation and the properties that will be used for finding the recurring patterns. This chapter is mainly adapted from Franziska Meier's master thesis [52].

## *A.1 Definition of Suffix Tree*

The suffix tree for a string $S$ of length $n$ is defined as a tree with the following properties:

1. the paths from the root to the leaf nodes have a one-to-one matching relationship with the suffixes of $S$,
2. each edge is labeled with a non-empty substring,
3. any internal node (except the root) has at least two children, and
4. no two edges out of a node can have labels that begin with the same character.

A string $S$ of length $n$ has $n$ suffixes, which means the suffix tree shall have $n$ leaves. Such a tree does not exist for all strings. Consider the string `banana`, its suffixes are listed in Table 13. The suffix `na` is a prefix of `nana`. Therefore, the path for suffix `na` won't end at a leaf node, which violates property 1 of a suffix tree. A simple solution is to pad $S$ with a special termination character not seen in the string

**Table 13:** Suffixes of string `banana`.

| suffixes | suffixes padded with $ |
|---|---|
| banana | banana$ |
| anana | anana$ |
| nana | nana$ |
| ana | ana$ |
| na | na$ |
| a | a$ |

(usually denoted $). Therefore each suffix will end with $, and none of them will be a prefix of another suffix.

## *A.2 Construction of a Suffix Tree*

The suffix tree construction method proposed by McCreight [50] inserts the suffixes of a string successively to an empty tree $T$, starting with the longest suffix. The insertion procedure is described in Algorithm 2. Figure 83 shows the corresponding process of constructing the suffix tree for string `banana`. When inserting substring `ana$` (Figure 83(e)), the existing edge labeled with `anana$` is split up at the end of `ana`, which is the longest prefix shared by `anana$` and `ana$`. A new node is then inserted and a new branch labeled with $ is created. The suffixes `na$` and `a$` are inserted in the same manner.

## *A.3 Pattern Extraction*

Once a suffix tree is built, we can traverse the tree to extract recurring patterns of any length, along with their number of occurrences and the location of each occurrence. A *pattern* is defined as the concatenated labels of a path from the root to any node in a suffix tree. For example, `ana` is a pattern in the suffix tree shown in Figure 83(g). A leaf node defines a suffix pattern, since the path from the root to the leaf is a suffix (Property 1).

The recurring patterns are extracted by traversing the tree, based on the following two facts about a suffix tree:

**Algorithm 2** McCreight's Suffix Tree Algorithm.

Input: string $S$ of length $n$
Output: a suffix tree rooted at $T$ for $S$.

1. The suffix tree is initialized with a single node, the root node $T$.
2. Insert string $S$ into the tree $T$. As a result, $T$ has one edge labeled with string $S$.
3. Insert suffixes $S[2, \ldots, n]$, $S[3, \ldots, n]$, $\ldots$, $S[n]$ successively into $T$ according to the following rules:
   - Check whether an edge leaving the root exists such that its label starts with the same character as the suffix to be inserted;
   - If such an edge exists, follow that edge until a mismatch between the label and the current suffix occurs. Insert a node at that point and create a new branch labeled with the rest of the suffix.
   - Otherwise add a new edge to the root labeled with the current suffix.
4. Repeat Step 3 until all the suffixes are inserted into $T$.

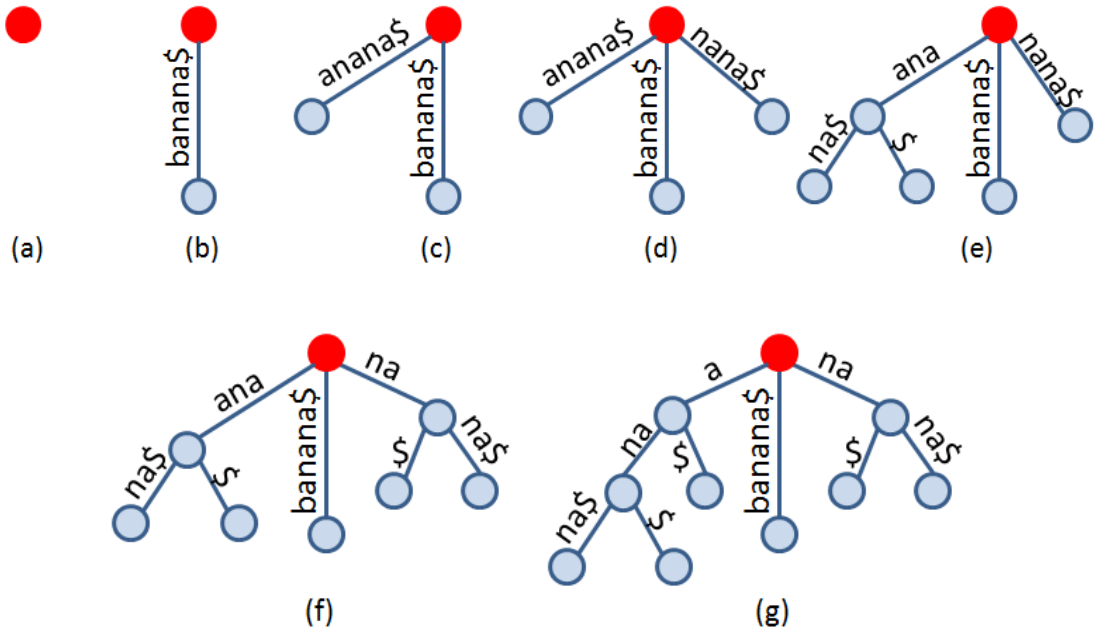



**Figure 83:** Build a suffix tree for string `banana`.

1. The number of occurrences of a pattern is equivalent to the number of leaf nodes of the subtree of that pattern.

2. A suffix pattern's starting position in the original string $S$ is determined by its length.

Fact 1 can be easily verified according to Algorithm 2. Fact 2 is derived from the definition of a suffix. Denote the length of a suffix pattern as $length(pattern)$. Suppose the string $S$ starts at position 1, the starting position of this pattern is:

$$\text{startIDX} = length(S) - length(pattern) + 1 \tag{18}$$

A pattern that ends at an internal node is the prefix of the suffix patterns that share the same path from the root to the internal node. Consequently, it appears more than once in $S$, so it has multiple starting positions. Each starting position is the same as that of a suffix pattern on that path. In other words, the starting positions of a pattern ending at an internal node are the union of the starting positions of the suffix patterns in the subtree rooted at that internal node. In Figure 83(g), pattern `na` appears twice in string `banana`: one starting at 3 and the other starting at 5. All these information can be gathered by a depth-first traversal of the suffix tree.

# REFERENCES


[1] *Advice from experts – how do you play pattycake with baby?* http://www.fisher-price.com/fp.aspx?st=10&e=expertadvice&content=30959.

[2] "Autism and family home movies: preliminary findings," *Journal of Autism and Developmental Disorders*, vol. 21, no. 1, pp. 43–49, 1991.

[3] "Autism: the phenotype in relatives," *Journal of Autism and Developmental Disorders*, vol. 28, no. 5, pp. 369–392, 1998.

[4] "What are infant siblings teaching us about autism in infancy?," *Autism Research*, vol. 2, no. 3, pp. 125–137, 2009.

[5] Adamson, L. B., Bakeman, R., Smith, C. B., and Walters, A. S., "Adults' interpretation of infants' acts," *Developmental Psychology*, vol. 23, no. 3, pp. 383–387, 1987.

[6] Ali, S., Basharat, A., and Shah, M., "Chaotic invariants for human action recognition," in *International Conference on Computer Vision (ICCV)*, 2007.

[7] Baranek, G. T., "Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9-12 months of age," *Journal of Autism and Developmental Disorders*, vol. 29, no. 3, pp. 213–224, 1999.

[8] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R., "Actions as space-time shapes," in *International Conference on Computer Vision (ICCV)*, 2005.

[9] Boiman, O. and Irani, M., "Detecting irregularities in images and in video," in *International Conference on Computer Vision (ICCV)*, 2005.

[10] Brand, M., "Physics-based visual understanding," *Computer Vision and Image Understanding*, vol. 65, no. 2, 1997.

[11] Bruner, J., "The social context of language acquisition," *Language and Communication*, vol. 1, no. 2/3, pp. 155–178, 1981.

[12] Bruner, J. and Sherwood, V., "Peekaboo and the learning of rule structures," *Early Interaction Play*, vol. 27, pp. 277–285, 1976.

[13] Burford, B., Kerr, A. M., and Macleod, H. A., "Nurse recognition of early deviation in development in home videos of infants with rett disorder," *Journal of Intellectual Disability Research*, vol. 47, no. 8, pp. 588–596, 2003.

[14] Cham, T.-J. and Rehg, J. M., "A multiple hypothesis approach to figure tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999.

[15] Chang, C.-C. and Lin, C.-J., *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[16] Colgan, S. E., Lanter, E., McComish, C., Watson, L. R., Crais, E. R., and Baranek, G. T., "Analysis of social interaction gestures in infants with autism," *Child Neuropsychology*, vol. 12, no. 4-5, pp. 307–319, 2006.

[17] Crais, E., Douglas, D. D., and Campbell, C. C., "The intersection of the development of gestures and intentionality," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 3, pp. 678–694, 2004.

[18] Cutler, R. and Davis, L. S., "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.

[19] Davis, J. and Bobick, A., "The representation and recognition of action using temporal templates," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 928–934, 1997.

[20] Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., and Varley, J., "Randomized, controlled trial of an intervention for toddlers with autism: The early start denver model," *Pediatrics*, vol. 125, no. 1, pp. e17–e23, 2010.

[21] Dollér, P., Rabaud, V., Cottrell, G., and Belongie, S., "Behavior recognition via sparse spatio-temporal features," in *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005.

[22] "Workshop on evaluation of articulated human motion and pose estimation, http://www.cs.brown.edu/ ls/ehum2/."

[23] Fan, Q., Bobbitt, R., Zhai, Y., Yanagawa, A., Pankanti, S., and Hampapur, A., "Recognition of repetitive sequential human activity," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] Fanti, C., *Towards Automatic Discovery of Human Movemes*. PhD thesis, California Institute of Technology, 2008.

[25] Feldman, R., "Parent-infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions," *J. Child Psychol. Psychiatry*, vol. 48, no. 3-4, pp. 329–354, 2007.

[26] Fiese, B., Poehlmann, J., Irwin, M., Gordon, M., and Curry-Bleggi, E., "A pediatric screening instrument to detect problematic infant-parent interactions: Initial reliability and validity in a sample of high- and low-risk infants," *Infant Mental Health Journal*, vol. 22, no. 4, pp. 463–478, 2001.

[27] Fogel, A., Nelson-Goens, G. C., and Hsu, H.-C., "Do different infant smiles reflect different positive emotions?," *Social Development*, vol. 9, no. 4, pp. 497–520, 2000.

[28] Garvey, C., "Some properties of social play," in *Play — its role in development and evolution* (Bruner, J., Jolly, A., and Sylva, K., eds.), pp. 570–583, Middlesex: Penguin, 1976.

[29] Garvey, C., *Play*. Cambridge: Havard University Press, 1977.

[30] Geweke, J., "Measurement of linear dependence and feedback between multiple time series," *Journal of American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.

[31] Gupta, A., Srinivasan, P., Shi, J., and Davis, L. S., "Understanding videos, constructing plots - Learning a visually grounded storyline model from annotated videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[32] Gustafson, G. E., Green, J. A., and West, M. J., "The infant's changing role in mother-infant games: the growth of social skills," *Infant Behavior and Development*, vol. 2, pp. 301–308, 1979.

[33] Hamid, R., Maddi, S., Bobick, A., and Essa, I., "Structure from statistics - unsupervised activity analysis using suffix trees," in *International Conference on Computer Vision (ICCV)*, 2007.

[34] Hay, D., Ross, H., and Goldman, B., "Social games in infancy," in *Play and learning* (Sutton-Smith, B., ed.), pp. 83–107, New York: Gardner Press, 1979.

[35] Hodapp, R. M. and Goldfield, E. C., "The use of mother-infant games as therapy with delayed children," *Early Child Development and Care*, vol. 13, no. 1, pp. 17–32, 1983.

[36] Hoey, J. and Little, J. J., "Value-directed human behavior analysis from video using partially observable markov decision processes," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1118–1132, 2007.

[37] Jhuang, H., Serre, T., Wolf, L., and Poggio, T., "A biologically inspired system for action recognition," in *International Conference on Computer Vision (ICCV)*, 2007.

[38] Junejo, I., Dexter, E., Laptev, I., and Pérez, P., "View-independent action recognition from temporal self-similarities," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, March 2010.

[39] Ke, Y., Sukthankar, R., and Hebert, M., "Event detection in crowded videos," in *International Conference on Computer Vision (ICCV)*, 2007.

[40] Laptev, I., "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.

[41] Laptev, I., Belongie, S. J., Perez, P., and Wills, J., "Periodic motion detection and segmentation via approximate sequence alignment," in *International Conference on Computer Vision (ICCV)*, pp. 816–823, 2005.

[42] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B., "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[43] Laptev, I. and Perez, P., "Retrieving actions in movies," in *International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[44] Laxton, B., Lim, J., and Kriegman, D., "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video," in *International Conference on Computer Vision (ICCV)*, 2007.

[45] Liu, J., Luo, J., and Shah, M., "Recognizing realistic actions from videos "in the wild"," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[46] Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., and Schopler, E., "Autism diagnostic observation schedule: a standardized observation of communicative and social behavior," *Journal of Autism and Developmental Disorders*, vol. 19, no. 2, pp. 185–212, 1989.

[47] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[48] Loy, C. C., Xiang, T., and Gong, S., "Modelling activity global temporal dependencies using time delayed probabilistic graphical model," in *International Conference on Computer Vision (ICCV)*, 2009.

[49] Mann, R., Jepson, A., and Siskind, J. M., "The computational perception of scene dynamics," *Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 113–128, 1997.

[50] McCreight, E., "A space-economical suffix tree construction algorithm," *Journal of the ACM*, vol. 23, pp. 262–272, 1976.

[51] McEvoy, R., Rogers, S., and Pennington, R., "Executive function and social communication deficits in young autistic children," *Journal of Child Psychology and Psychiatry*, vol. 34, pp. 563–578, 1993.

[52] Meier, F., "Automatic segmentation of human activities into motion subtasks," Master's thesis, Technical University Munich, 2008.

[53] Minnen, D., Essa, I., Isbell, C., and Starner, T., "Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery," in *IEEE International Conference on Data Mining (ICDM)*, 2007.

[54] Moore, D., Essa, I., and Hayes, M., "Exploiting human actions and object context for recognition tasks," in *International Conference on Computer Vision (ICCV)*, pp. 80–86, 1999.

[55] Mundy, P., Sigman, M., and Kasari, C., "A longitudinal study of joint attention and language development in autistic children," *Journal of Autism and Developmental Disorders*, vol. 20, pp. 115–128.

[56] Munguia-Tapia, E., Choudhury, T., , and Philipose, M., "Building reliable activity models using hierarchical shrinkage and mined ontology," in *Proceedings of Pervasive*, 2006.

[57] Needham, C., Santos, P., Magee, D., Devin, V., Hogg, D., and Cohn, A., "Protocols from perceptual observations," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 103–136, 2005.

[58] Niebles, J. C., Wang, H., and Fei-Fei, L., "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[59] Nowozin, S., Bakir, G., and Tsuda, K., "Discriminative subsequence mining for action classification," in *International Conference on Computer Vision (ICCV)*, 2007.

[60] Odom, S., Rogers, S., McDougle, C., Hume, K., and McGee, G., "Early intervention for children with autism spectrum disorder," in *Handbook of Developmental Disabilities* (Odom, S., Horner, R., and Snell, M., eds.), pp. 199–223, Guildford Press, 2009.

[61] Oliver, N., Horvitz, E., and Garg, A., "Layered representations for human activity recognition," in *Proceedings of IEEE International Conference on Multimodal Interfaces*, pp. 3–8, 2002.

[62] Oliver, N., Rosario, B., and Pentland, A., "Statistical modeling of human interactions," in *IEEE CVPR Workshop on the Interpretation of Visual Motion*, 1998.

[63] Park, A. S. and Glass, J. R., "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Lauguage Processing*, vol. 16, pp. 186–197, January 2008.

[64] Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J., "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[65] Pinhanez, C. and Bobick, A., "Pnf propagation and the detection of actions described by temporal intervals," in *DARPA Image Understanding Workshop*, 1997.

[66] Polana, R. and Nelson, R., "Detection and recognition of periodic, non-rigid motion," *International Journal of Computer Vision*, vol. 23, no. 3, pp. 261–282, 1997.

[67] Prabhakar, K., Oh, S., Wang, P., Abowd, G. D., and Rehg, J. M., "Temporal causality for the analysis of visual events," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[68] Rabiner, L., "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[69] Ramanan, D. and Forsyth, D., "Automatic annotation of everyday movements," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[70] Rochat, P., Querido, J. G., and Striano, T., "Emerging sensitivity to the timing and structure of protoconversation in early infancy," *Developmental Psychology*, vol. 35, no. 4, pp. 950–957, 1999.

[71] Rogers, S. and Vismara, L., "Evidence-based comprehensive treatments for early autism," *Journal of Clinical Child and Adolescent Psychology*, vol. 37, no. 1, pp. 8–38, 2008.

[72] Rowley, H. A., Baluja, S., and Kanade, T., "Neural network-based face detection," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 23–38, January 1998.

[73] Sacks, H., Schegloff, E., and Jefferson, G., "A simplest systematics for the organization of turn-taking in conversation," *Lauguage*, vol. 50, pp. 696–735, 1974.

[74] Scharstein, D. and Szeliski, R., "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[75] Schuldt, C., Laptev, I., and Caputo, B., "Recognizing human actions: A local SVM approach," in *Proceedings of International Conference on Pattern Recognition*, pp. 32–36, 2004.

[76] Shechtman, E. and Irani, M., "Space-time behavior based correlation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[77] Shechtman, E. and Irani, M., "Matching local self-similarities across image and videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[78] Shi, Y., Bobick, A., and Essa, I., "Learning temporal sequence model from partially labeled data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[79] Sidner, C. and Lee, C., "Attentional gestures in dialogues between people and robots," in *Engineering Approaches to Conversational Informatics* (Nishida, T., ed.), Wiley and Sons, 2007.

[80] Sminchisescu, C. and Triggs, B., "Kinematic jump processes for monocular 3d human tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 69–76, 2003.

[81] Smith, T., Groen, A., and Wynn, J., "Randomized trial of intensive early intervention for children with pervasive developmental disorder," *American Journal of Mental Retardation*, vol. 105, no. 4, pp. 269–285, 2000.

[82] Song, Y., Goncalves, L., and Perona, P., "Unsupervised learning of human motion," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 25, pp. 1–14, 2003.

[83] Starner, T., Weaver, J., and Pentland, A., "Real-time American sign language recognition using desk and wearable computer-based video," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[84] "Trec video retrieval evaluation, http://www-nlpir.nist.gov/projects/trecvid/."

[85] Ukkonen, E., "Approximate string-matching over suffix trees," in *CPM '93: Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching*, pp. 228–242, 1993.

[86] Wang, P., Abowd, G. D., and Rehg, J. M., "Quasi-periodic event analysis for social game retrieval," in *International Conference on Computer Vision (ICCV)*, 2009.

[87] Wang, P. and Rehg, J. M., "A modular approach to the analysis and evaluations of particle filters for figure tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[88] Weiner, P., "Linear pattern matching algorithms," in *Proceedings of IEEE Symposium on Switching and Automata Theory*, pp. 1–11, 1973.

[89] Werner, E. and Dawson, G., "Regression in autism: Validation of the phenomenon using home videotapes," *Arch Gen Psychiatry*, vol. 62, pp. 889–895, 2005.

[90] Werner, E., Dawson, G., Osterling, J., and Dinno, N., "Brief report: recognition of autism spectrum disorder before one year of age: a retrospective study based on home videotapes," *Journal of Autism and Developmental Disorders*, vol. 30, no. 2, pp. 157–162, 2000.

[91] Wetherby, A. and Prutting, C., "Profiles of communicative and cognitive-social abilities in autistic children," *Journal of Speech and Hearing Research*, vol. 27, pp. 364–377, 1984.

[92] Wilson, A., Bobick, A., and Cassell, J., "Recovering the temporal structure of natural gesture," in *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 1996.

[93] Wilson, A. D. and Bobick, A. F., "Parametric hidden markov models for gesture recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, 1999.

[94] Wong, S.-F., Kim, T.-K., and Cipolla, R., "Learning motion categories using both semantic and structural information," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[95] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., and Rehg, J., "A scalable approach to activity recognition based on object use," in *International Conference on Computer Vision (ICCV)*, 2007.

[96] Wyatt, D., Philipose, M., and Choudhury, T., "Unsupervised activity recognition using automatically mined common sense," in *Proceedings of AAAI*, pp. 21–27, 2005.

[97] Yacoob, Y. and Black, M., "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, no. 2, pp. 232–247, 1999.

[98] Yang, J., Wang, W., and Yu, P. S., "Infominer: mining surprising periodic patterns," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 395–400, ACM Press, 2001.

[99] Yilmaz, A. and Shah, M., "Actions as objects: A novel action representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[100] Yuan, J., Meng, J., Wu, Y., and Luo, J., "Mining recurring events through forest growing," *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, vol. 18, no. 11, pp. 1597–1607, 2008.

[101] Yuen, J., Russell, B., Liu, C., and Torralba, A., "Labelme video: Building a video database with human annotations," in *International Conference on Computer Vision (ICCV)*, 2009.

[102] Zhai, Y. and Shah, M., "Visual attention detection in video sequences using spatiotemporal cues," in *ACM International Conference on Multimedia*, 2006.

[103] Zhong, H., Shi, J., and Visontai, M., "Detecting unusual activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[104] Zhou, Y., Yan, S., and Huang, T. S., "Pair-activity classification by bi-trajectories analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.